

NAVAIR Delivery Order 1293

Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative

Final Report

Contract Number: N66032-94-D-0012

Prepared For:

Mr. James Leach
Attn: AMSAM-CIC-T-E
Building 4722
Redstone Arsenal, AL 35894
Telephone: 256-842-7389
Fax: 256-842-7360
Email: jimmy.leach@redstone.army.mil

and

Mr. M. K. Miles
Attn: CECW-EE
20 Massachusetts Ave. NW
Washington, D.C. 20314-1000
Telephone: 202-761-8885
Fax: 202-761-4002
Email: moody.k.miles@usace.army.mil

Prepared By:

Intergraph Solutions Group
170 Graphics Drive
Madison, AL 35878



INTERGRAPH

SOLUTIONS GROUP

This document includes data that shall not be disclosed outside the Government. However, the Government shall have the right to duplicate, use, or disclose the data to the extent provided in NAVAIR Delivery Order 1293. This restriction does not limit the Government's right to use information contained in these data if they are obtained from another source without restriction.

Table of Contents

1	INTRODUCTION	1
2	PURPOSE	2
3	BENEFITS	3
3.1	DESIGN	4
3.2	STORAGE.....	4
3.3	RETRIEVAL	5
3.4	REPRODUCTION.....	7
3.5	ADDITIONAL APPLICATIONS OF DOCUMENT AND CONTENT CAPTURE.....	7
3.5.1	<i>Support for Future Systems.....</i>	7
3.5.2	<i>Disposition Program.....</i>	9
3.5.3	<i>Knowledge Management.....</i>	9
3.5.4	<i>Document Workflow Management.....</i>	10
4	USACE CURRENT SITUATION	12
4.1	SITE VISIT DATA	12
4.1.1	<i>Ft. Worth District Notes</i>	13
4.1.2	<i>Mobile District Notes</i>	14
4.1.3	<i>Seattle District Notes.....</i>	16
4.2	SITE VISIT OBSERVATIONS	17
4.3	CONVERSION CHALLENGES.....	21
4.3.1	<i>Paper.....</i>	21
4.3.2	<i>Deterioration.....</i>	22
4.3.3	<i>Volume</i>	24
4.3.4	<i>Standalone Storage</i>	26
4.3.5	<i>Legacy Electronic Formats</i>	27
4.3.6	<i>Metadata Collection.....</i>	28
4.3.7	<i>Information Reuse.....</i>	30
4.3.8	<i>Funding.....</i>	31
5	USACE RECOMMENDATIONS.....	32
5.1	CONVERSION STRATEGY	32
5.1.1	<i>Strategy Goals.....</i>	32
5.1.2	<i>Document Conversion Activity.....</i>	33
5.1.3	<i>Business Case Review Activity.....</i>	34
5.1.4	<i>Other Issues.....</i>	37
5.2	DOCUMENT CAPTURE	40
5.2.1	<i>Document and Content Capture Defined</i>	41
5.2.2	<i>Vectorization</i>	43
5.2.3	<i>Text Recognition.....</i>	44
5.3	METADATA CAPTURE	46
5.4	DOCUMENT REPOSITORY	47
5.5	DOCUMENT ACCESS.....	48
5.5.1	<i>Document Management</i>	48
5.5.2	<i>Planning for Digital Permanence</i>	49
5.6	BULK LOADING.....	50
5.7	DCRDS.....	51
5.8	CONVERSION PROCEDURE.....	53
6	ADDITIONAL CONSIDERATIONS.....	55
6.1	RETURN ON INVESTMENT/TOTAL COST OF OWNERSHIP.....	55
6.1.1	<i>Direct Benefits.....</i>	55
6.1.2	<i>Indirect Benefits</i>	59

6.1.3	Purchasing Considerations	60
6.2	APPLICABLE DOCUMENTS	62
6.3	EDMS USE WITHIN DoD ORGANIZATIONS	64
6.4	FUTURE IMPROVEMENT	65
7	CONCLUSION	67
APPENDIX A.	GLOSSARY	A-1
APPENDIX B.	METADATA	B-1
APPENDIX C.	DOCUMENT TYPES	C-1
APPENDIX D.	1999 ADCS PROJECT SYNOPSIS	D-1
APPENDIX E.	SAMPLE CONVERSION WORKSHEET	E-1
APPENDIX F.	STANDARDS AND ESTIMATES	F-1
F.1	STANDARDS	F-1
F.1.1	Introduction	F-1
F.1.2	Final Adjustment Factor	F-2
F.1.3	Units of Measure (Digital)	F-3
F.1.4	Transmission Time Standards	F-4
F.1.5	Pages, File Cabinets, Boxes, and Linear Feet	F-4
F.1.6	Digital Media Capacity	F-5
F.1.7	Pixel Sizes and Pixels per Image	F-6
F.2	ESTIMATING STORAGE REQUIREMENTS FOR DIGITIZED DOCUMENT TYPES	F-10
F.2.1	Scanned Letter Size Pages	F-11
F.2.2	Scanned Engineering Drawings / Large Format Documents	F-12
F.2.3	Scanned Microforms	F-13
F.2.4	Scanned Miscellaneous Documents	F-13
F.2.5	COLD / COOL	F-14
F.2.6	Digitized Multimedia Formats	F-14
F.2.7	Medical Records	F-14
F.2.8	Trees and Paper Requirements	F-15
APPENDIX G.	DRAFT INTERVIEW FORM	G-1
APPENDIX H.	SAMPLE DCRD	H-1
PREFACE	H-1
H.1	INTRODUCTION	H-1
H.1.1	Acronyms	H-1
H.1.2	Scope	H-2
H.1.3	Purpose/Conversion Level	H-2
H.1.4	Background	H-3
H.1.5	Interchangeable Terminology	H-3
H.1.6	Target Systems/ File Formats	H-4
H.1.7	File Format	H-5
H.1.8	Documentation	H-5
H.1.9	Hard Copy	H-5
H.1.10	Ownership	H-5
H.1.11	Additions/Revisions	H-6
H.2	DRAWING FILE ORGANIZATION	H-7
H.2.1	Available Drawing Area	H-7
H.2.2	File Accuracy (Units)	H-8
H.2.3	Origin (Global Origin)	H-8
H.2.4	Electronic Drawing File Naming Conventions	H-9
H.2.5	Industry Standard Model File, Naming Convention	H-10
H.2.6	Industry Standard Sheet File Naming Convention	H-12

H.2.7	PMIG Optional Model File Naming Convention	H-14
H.2.8	Optional Sheet File Naming Convention.....	H-16
H.2.9	Coordination Between Sheet File Name and Sheet Identifier	H-16
H.3	GRAPHIC CONTENT	H-16
H.3.1	Drawing Graphics.....	H-16
H.3.2	Line Widths	H-17
H.3.3	Line Types/Styles.....	H-18
H.3.4	Line Color	H-18
H.3.5	Screening (Halftoning).....	H-19
H.3.6	Text Styles/Fonts	H-19
H.3.7	Plotting.....	H-20
H.3.8	Border Sheets	H-22
H.3.9	Sheet Sizes	H-22
H.3.10	Title Block	H-22
H.3.11	Issue Block	H-23
H.3.12	Management Block.....	H-24
H.3.13	Project Identification Block/Sheet Title Block	H-24
H.3.14	Sheet Identification Block.....	H-24
H.3.15	Drawing Scales	H-25
H.3.16	Dimensioning in Metric (SI).....	H-27
H.3.17	Millimeters	H-27
H.3.18	Meters	H-28
H.3.19	Large Units of Measure	H-28
H.3.20	Dual Units	H-29
H.4	LEVEL/LAYER ASSIGNMENTS	H-29
H.4.1	Levels/Layers	H-29
H.4.2	Level/Layer Naming Conventions	H-31
H.4.3	ISO Format	H-32
H.4.4	Model Files	H-33
H.4.5	Level/Layer Assignment Tables.....	H-33
H.5	DELIVERABLES AND DATA EXCHANGE	H-34
H.5.1	Delivery Media.....	H-34
H.6	REFERENCES.....	H-35
	GENERAL DRAFTING LINES LIBRARY	H-55
	GENERAL DRAFTING OBJECTS LIBRARY	H-56
	GENERAL DRAFTING SYMBOLS LIBRARY	H-61
	UTILITIES SYMBOLS LIBRARY	H-63
	UTILITIES PATTERNS LIBRARY	H-65
	PLUMBING LINES LIBRARY	H-66
	FIRE SUPPRESSION LINES LIBRARY	H-67
	PLUMBING LINES LIBRARY	H-69
	PLUMBING SYMBOLS LIBRARY	H-74
	MECHANICAL SYMBOLS LIBRARY	H-81
	FIRE SUPPRESSION SYMBOLS LIBRARY	H-86

Table of Figures

Figure 3-1 – Document Storage Costs	5
Figure 3-2 – Lifecycle Records Management	8
Figure 5-1 – Recommended Document Capture and Access Strategy	40
Figure 5-2 – Character Recognition Technologies	45
Figure 5-3 – Recommended Converted Document Access Scheme	49
Figure 5-4 – Conversion Decision Matrix	53
Figure 6-1 – The Cost of Paper	55
Figure 6-2 – Cost Analysis of Various Revision Methods	57
Figure 6-3 – Lifetime Cost-Benefit Analysis	58
Figure D-1 – 1999 EDMS and ADCS Project Schedule	D-7
Figure F-1 – Creating an Image File from the Printed Page	F-7
Figure F-2 – Resolution and File Size Comparison	F-8
Figure H-1 – Available Drawing Size	H-7
Figure H-2 – Origins in MicroStation and AutoCAD	H-9
Figure H-3 – Sheet File Composition	H-10
Figure H-4 – Model File Naming Convention	H-11
Figure H-5 – Sheet Filing Name Convention	H-12
Figure H-6 – Typical File Structure	H-14
Figure H-7 – Optional Model File Naming Convention	H-14
Figure H-8 – Optional Sheet File Naming Convention	H-16
Figure H-9 – Typical Designer Identification Block	H-23
Figure H-10 – Issue Block	H-23
Figure H-11 – Management Block	H-24
Figure H-12 – Project Identification/Sheet Title Block	H-24
Figure H-13 – Sheet Identification Block	H-25
Figure H-14 – Dimension (mm)	H-27
Figure H-15 – Dimension (m)	H-28
Figure H-16 – Dimensions for Metric Measurements with Four or More Digits	H-28
Figure H-17 – Typical Levels/Layers Contained in a Sheet File	H-30
Figure H-18 – Sheet- and Model-Specific Information	H-31
Figure H-19 – Level/Layer Naming Format Methods	H-32
Figure H-20 – ISO 13567-2 Level/Layer Naming Method	H-33
Figure H-21 – General Lines	H-55
Figure H-22 – General Objects 1	H-56
Figure H-23 – General Objects 2	H-57
Figure H-24 – General Objects 3	H-58
Figure H-25 – General Objects 4	H-59
Figure H-26 – General Objects 5	H-60
Figure H-27 – General Symbols 1	H-61
Figure H-28 – General Symbols 2	H-62
Figure H-29 – Utilities Symbols 1	H-63
Figure H-30 – Utilities Symbols 2	H-64
Figure H-31 – Utilities Patterns	H-65
Figure H-32 – Architectural Plumbing Lines	H-66
Figure H-33 – Fire Suppression Lines	H-67
Figure H-34 – Fire Suppression Lines 2	H-68
Figure H-35 – Plumbing Lines 1	H-69
Figure H-36 – Plumbing Lines 2	H-70
Figure H-37 – Plumbing Lines 3	H-71
Figure H-38 – Plumbing Lines 4	H-72
Figure H-39 – Plumbing Lines 5	H-73
Figure H-40 – Plumbing Symbols 1	H-74
Figure H-41 – Plumbing Symbols 2	H-75
Figure H-42 – Plumbing Symbols 3	H-76

Figure H-43 – Plumbing Symbols 4	H-77
Figure H-44 – Plumbing Symbols 5	H-78
Figure H-45 – Plumbing Symbols 6	H-79
Figure H-46 – Plumbing Symbols 7	H-80
Figure H-47 – Mechanical Lines 1	H-81
Figure H-48 – Mechanical Lines 2	H-82
Figure H-49 – Mechanical Lines 3	H-83
Figure H-50 – Mechanical Lines 4	H-84
Figure H-51 – Mechanical Lines 5	H-85
Figure H-52 – Fire Suppression Symbols 1	H-86
Figure H-53 – Fire Suppression Symbols 2	H-87
Figure H-54 – Fire Suppression Symbols 3	H-88
Figure H-55 – Fire Suppression Symbols 4	H-89
Figure H-56 – Fire Suppression Symbols 5	H-90
Figure H-57 – Fire Suppression Symbols 6	H-91
Figure H-58 – Fire Suppression Symbols 7	H-92
Figure H-59 – Fire Suppression Symbols 8	H-93
Figure H-60 – Fire Suppression Symbols 9	H-94
Figure H-61 – Fire Suppression Symbols 10	H-95
Figure H-62 – Fire Suppression Symbols 11	H-96
Figure H-63 – Fire Suppression Symbols 12	H-97
Figure H-64 – Fire Suppression Symbols 13	H-98
Figure H-65 – Fire Suppression Symbols 14	H-99
Figure H-66 – Fire Suppression Symbols 15	H-100
Figure H-67 – Fire Suppression Symbols 16	H-101
Figure H-68 – Fire Suppression Symbols 17	H-102
Figure H-69 – Fire Suppression Symbols 18	H-103
Figure H-70 – Fire Suppression Symbols 19	H-104
Figure H-71 – Fire Suppression Symbols 20	H-105
Figure H-72 – Fire Suppression Symbols 21	H-106

Table of Tables

Table 3-1 – Benefits of Using Metadata	6
Table 4-1 – Ft. Worth Document Types and Quantities	13
Table 4-2 – Mobile Document Types and Quantities	14
Table 4-3 – Seattle Document Types and Quantities	16
Table 4-4 – 1999 Document Estimate per Field Site	25
Table 5-1 – Business Case Decision Table Requirements Determination	38
Table 6-1 – Cost-Benefit Comparison of Raster-Enabled Process	57
Table B-1 – Indexing Data Elements	B-1
Table B-2 – Non-Indexing and Table Data Elements	B-1
Table C-1 – Document Format Examples	C-2
Table C-2 – Proposed Document Types	C-2
Table D-1 – Current Document Estimate Per Field Site	D-5
Table D-2 – ADCS Project Total Cost Summary	D-7
Table D-3 – ADCS Project Total Cost Per FY	D-8
Table D-4 – ADCS Project Cost Per Site	D-8
Table F-1 – Storage Capacities of Standard Storage Media	F-5
Table H-1 – Acronyms	H-1
Table H-2 – Interchangeable Terminology	H-3
Table H-3 – MicroStation Working Units and Global Origins	H-8
Table H-4 – Discipline Codes/Designators	H-10
Table H-5 – Standard Drawing Type Codes	H-12
Table H-6 – DCRD Sheet Type Codes/Designators	H-13

Table H-7 – DCRD Optional Discipline Type Codes/Designators	H-15
Table H-8 – DCRD Optional Drawing Type Codes	H-15
Table H-9 – Comparison of Line Widths	H-17
Table H-10 – Standard Line Types/Styles	H-18
Table H-11 – Screen Color Comparison and Associated Line Widths	H-19
Table H-12 – Halftone Colors	H-20
Table H-13 – ISO, ANSI, Sheet Size Comparison	H-22
Table H-14 – Drawing Scales	H-25
Table H-15 – Inch-Pound Text Sizes	H-26
Table H-16 – Metric Text Sizes	H-27
Table H-17 – Color Comparison and Associated Line Widths	H-37
Table H-18 – Mechanical, Details	H-44
Table H-19 – Mechanical, HVAC Plans	H-45
Table H-20 – Mechanical, Machine Design	H-47
Table H-21 – Plumbing, Demolition Plan	H-48
Table H-22 – General	H-49
Table H-23 – Fire Protective/Suppression	H-50
Table H-24 – HTRW/Environmental	H-51
Table H-25 – Mechanical	H-52
Table H-26 – Plumbing	H-53
Table H-27 – Utilities	H-54

1 Introduction

This document provides an overview of a strategy or methodology for planning and implementing automated document conversion in order to reduce the overhead labor and cost associated with storing documents. If implemented, this strategy offers faster response, greater reliability, and lower cost in the retrieval process.

Access to information is of primary importance to the United States Army Corps of Engineers' (Corps, USACE) mission and workflow. Lengthy searches for data and documents impair the Corps' mission and workflow and add costly delays. Inadvertent loss of data and documents, whether due to misplacement or deterioration, multiply these delays. The Automated Document Conversion Strategy (ADCS) initiative is intended to leverage document conversion technologies and methods to provide significant improvements in the way Corps workers access and use legacy information to support present and future projects.

The Corps has identified overwhelming and urgent needs to reduce significant storage costs, increase productivity, shorten design and drafting time, shorten information search and retrieval time, decrease repetition of effort, increase consistency with past work, and preserve the Corps' extensive, irreplaceable knowledge assets before they are lost forever. To achieve these goals, the Corps needs a methodology that provides a standardized, efficient process that ensures the documents are fully captured, converted, and indexed so they can be easily located, reproduced, and reused after conversion.

A total document conversion solution has many interdependent elements. No single commercial, off-the-shelf (COTS) solution will alleviate the increasing problems the Corps has with documents. To succeed with document conversion, the Corps must adopt a process or methodology to accomplish meaningful document capture while at the same time providing user with the means of searching online repositories or archives for information needed to accomplish their work objectives. The most important part of this methodology is the capture of information describing the documents' content. Once collected, this descriptive information, or "metadata," provides a searchable index that tremendously increases the converted documents' value. Metadata can also be leveraged when implementing document or records management; the index can be bulk-loaded into the new system to immediately upgrade the document archive to a future document or records management environment.

Access to the Corps' library of valuable data continues to be a driving force in providing Corps services and performing Corps functions. Without easy accessibility to existing information, workers suffer costly delays in getting the information they need to perform their jobs. This document describes proven methods for data conversion and metadata collection for engineering drawings and other Corps knowledge assets. The tools and processes outlined in this document can help the Corps achieve its goals for document conversion. Through planning and consistent application of an informed, practical document capture plan, information once housed in remote file cabinets can be made available in reusable electronic format throughout the Corps. Collection sites and warehouses will decrease in size, reducing storage cost. And improvements in document search and retrieval procedures will reduce time-associated costs in production.

2 Purpose

Corps workers have been generating valuable, information-rich documents since the Corps' inception, with computerization progressively increasing the speed of information flow for at least the last 15 years. This extensive library of documentation is currently distributed across a wide variety of recordkeeping systems and housed in a variety of media – everything from personal hard drives, to server directories, to paper, film, and other physical formats. In addition to statutory reasons for protecting and preserving Corps documentation, the Corps has an ongoing need to locate and reuse information as well as evaluate its importance to the organization. Without an organized, cohesive system of information storage, reference, and retrieval, the Corps has no easy way of determining what information it has; when, where, and how it can be accessed; and how a user can differentiate important information from unimportant information.

The Corps identified the following goals to be addressed by the ADCS initiative in particular, and document conversion in general:

- Digitize data formatted non-electronically, and transform the common operating environment to an all-digital one.
- Use existing data conversion facilities to limit and reduce the total infrastructure requirement for the Department of Defense (DoD).
- Collect, index, and integrate legacy documents and information into a future Distributed Knowledge Environment (DKE).
- Develop user-based search and retrieval tools based on user requirements and the DKE.
- Improve on the DoD's ability to reduce acquisition/production lead time related to the conversion of input data and in making information accessible to users.
- Reduce Freedom of Information Act (FOIA) request cost to the customer to better serve the public.

Using the guidelines established by the Corps, this document provides assistance toward making the Corps' extensive library of historical information available in a useful digital format, with the minimum digital format being raster information. Digital data must support all stakeholders in the Corps. These stakeholders include two distinct communities: the authors and managers of technical documentation who are responsible for ensuring that technical data is current and can be referenced properly; and the users of the documentation who leverage existing content to effect repair, maintenance, or acquisition actions in support of a particular project.

This document sets forth tools for building the strategy that will enable the Corps to fulfill its goals for converting legacy documents workers require to do their jobs now and in the future, drastically reducing wasted hours spent searching for existing information. This document also addresses specific considerations for building conversion requirements for engineering drawings and other technical documentation.

3 Benefits

Sharing documents is one of the most frequent forms of collaboration within any enterprise. Once committed to electronic file formats, documents may be shared through Web pages, transmitted via e-mail, made available through file servers, and integrated into document portals.

Converting hard-copy documents and drawings is expected to significantly reduce the cost of doing business throughout the Corps, as well as reduce the cost of design tasks, file storage, information retrieval, and document reproduction. Conversion of existing records to scanned images accomplishes several additional objectives:

- Storage space and access speed concerns are greatly alleviated, as servers can store literally millions of images in a small amount of space for a fraction of the cost of mere real estate, and a server can provide instant access to any image given proper metadata capture.
- For fragile or deteriorating documents, of which Corps Districts have an ever-increasing number, scanned images provide access to information without subjecting the originals to further physical handling.
- The Corps will gain the ability to create a complete backup of all electronic images and their metadata for storage at an offsite facility in the event of a disaster affecting the document archives.

The strategy defined in this report enables the Corps to realize all these benefits. First, it provides the Corps with the tools to build a complete process for converting documents or drawings to reusable digital media, capturing metadata (attributes) about the documents or drawings, storing documents or drawings in a permanent digital storage folder or archive, and finally providing access to these converted documents/drawings via a user-friendly Web interface while at the same time providing a means of bulk loading that same data into future systems such as a document management system, records management system, or some other customized information management system.

The Corps is quickly approaching the point where the volume of its data will exceed the Corps' ability to manage it. The cost of accessing that data increases continually without proper management. Document capture is the key to making content more accessible, and therefore more valuable. The process of document and content capture can be broken down into three steps:

- Capturing inputs, both paper and digital
- Processing the input to capture its metadata
- Releasing the result into awaiting repositories or applications

Each step adds value to the document. An electronic document on its own, even without metadata or a process by which to re-use its content, has increased in value over the original paper because it can be infinitely copied, reproduced, and transmitted without any risk of damaging the original. But a document that has been captured, converted, indexed, and validated

can have a significantly higher value, depending on the extent and usefulness of the process or repository application being served.

3.1 Design

Access to legacy data continues to be a driving force in providing Corps services and performing Corps functions. A large quantity of the design work performed at the Corps consists of maintenance or additions to existing structures or property. Frequently, this design work utilizes existing design data from original or previous construction, which must be laboriously recreated from deteriorating hard-copy formats. Significant cost savings will be achieved if these hard-copy documents are converted to electronic documents. Once converted, electronic documents can be more easily used with new designs and the electronic files can be transferred to contractors electronically.

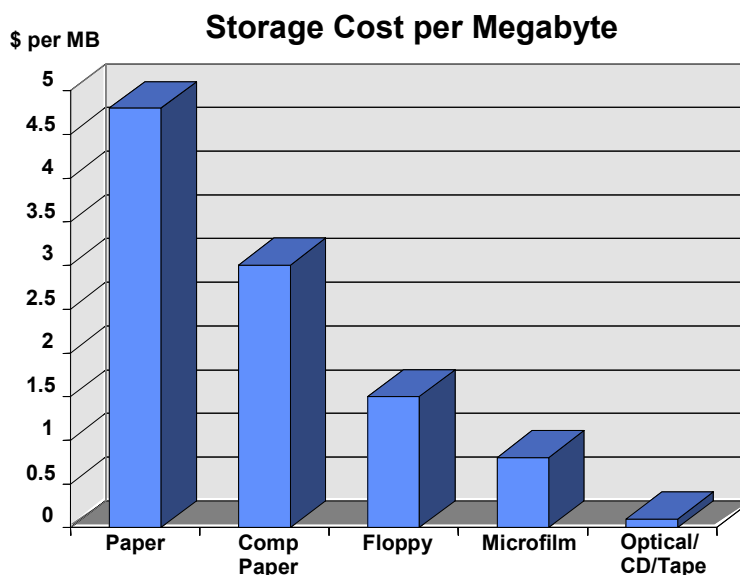
The need to capture, modify, and distribute existing paper designs within the environment of today's computing technology is steadily increasing. Considering that each successive stage in a product development cycle – including design, production, and support services – uses substantially more documentation than its predecessor, the benefits of integrating existing information into present and future work processes grows exponentially. With the implementation of a managed electronic archive, searching time, as well as re-engineering time, become past history. This translates into real dollar savings. Substantial savings can also be realized through a managed revision process. The reliance on manual drafting and control of drawing revisions on older documentation can be put to rest.

3.2 Storage

Throughout the Corps, organizations store millions of documents of various sizes. These documents must be stored on site (in most cases) to keep them readily accessible. Every year, the Corps pays for the floor space to house paper documents. Although an investigation of the total Corps-wide cost of storage space (real estate, upkeep, associated labor, etc.) has not been performed, USACE Headquarters estimated a savings of about one million dollars per year if the need for storing hard-copy documents were eliminated. (Headquarters is one of 63 Corps facilities and its approximately 1,000 employees are only a fraction of the estimated 35,000 Corps employees nationwide.)

In the last ten years, the cost of electronic storage and retrieval has decreased. Over the same period of time, the cost of storing and retrieving paper has increased. Facilities are more expensive, labor is more expensive (and facing reductions), paper databases are becoming larger and more cumbersome, and customer service expectations are increasing. While the cost per megabyte of electronic storage continues to decline, industry estimates of paper storage costs are rising higher than 25–30 cents per sheet stored, counting real estate lease and rental costs, overhead labor costs, and other hidden charges most organizations pay reflexively. The following figure compares the cost of various storage forms against paper, in dollars per megabyte.

Figure 3-1 – Document Storage Costs



CD-ROM and digital tape formats are the least expensive storage media by a wide margin, as shown in Figure 3-1 above (CD-ROM: less than \$0.002 per megabyte; DLT-IV: less than \$0.0035 per megabyte). Reducing or eliminating hard copy document storage stands to significantly improve Corps finances.

3.3 Retrieval

Retrieving documents for review or reproduction in the existing Corps environment can be both tedious and time consuming. Often, files are difficult to find, or access to the documents is limited. In some cases, documents are stored in large file rooms in remote areas within Corps facilities. The Corps estimated employees require up to 30 minutes to retrieve a single document, provided the document was not actually stored offsite at another locations or at another Corps office. Site visit data indicated documents sometimes take hours to retrieve. Offsite documents can take days to retrieve. These delays impose costs, but converting documents into electronic files will reduce or eliminate these delays by making the files available to users' computer workstations.

Most organizations are guilty of duplicating information they already possess simply because the information they require lies buried, poorly registered or indexed, in a filing cabinet somewhere in the organization. This phenomenon was once summarized with the observation, "if we only knew what we knew, we'd be four times as productive." Industry estimates indicate that 3 to 5 percent of an organization's documents are misfiled at any one time, and that more than 7 percent of an executive's time is wasted looking for misfiled documents. Often it becomes necessary to recreate the document: a time-consuming effort, costing organizations substantial amounts in what are, ultimately, avoidable labor hours.

There are physical reasons documents might be unavailable. In the Corps, documents continue to suffer the ravages of time – mildew, dampness, fading ink, and damaged paper are only the most

common symptoms of age. Aging film archives are subject to chemical self-destruction. Microfilm archives, aperture cards, and other microforms are subject to these risks as well as the risks of poor transcription quality and viewer obsolescence or mechanical breakdown. A document may even be merely lost – borrowed and never returned, misfiled, or misplaced.

The most valuable attributes of any document are the content and context in which it was created. These attributes are largely lost once active paper or microfilm documents are committed to archives, and the information ceases to have much value unless someone knows that it exists and where it is stored. In an electronic archive, metadata, or “information about information,” is a tool for capturing these valuable attributes and storing them with the document. Classifying newly created documents – conventionally seen as a “basement” activity that takes place after the documents have moved beyond regular access – is of paramount importance in keeping a document retrievable and maintaining its value. Addressing the requirement to collect and store metadata is best recognized as a “front office” task.

Table 3-1 – Benefits of Using Metadata

Time-tested and reliable	Librarians experimented and fine-tuned this methodology. Metadata has become the preferred method for multi-national organizations and archival facilities. Metadata implementation follows universal bibliographic standards established for the card catalog and developed by technology pioneers.
No database pointer records	Using metadata is similar to querying a database using document descriptors. But unlike a database which uses pointer records to direct you to document location, metadata lives inside a document and remains there no matter where the document moves.
Support for all media formats and document types	Metadata allows the user to catalog all types of digital resources. Technology facilitates easy creation of multimedia files, such as audio and video formats. Accessibility laws mandate that organizations use them to assist people with disabilities. Metadata makes these file formats part of the mainstream document collection. Metadata also search-enables all file formats, including unique formats like PDF.
Cost-effectiveness and performance	An automated system augments manual metadata implementation. Metadata systems require minimal user programming, and schedule updates for growing document collections. High-speed document retrieval is measured in milliseconds.
Integration with existing methods	Metadata works with existing archival procedures, search engines, and search portals. In addition, an automated metadata system can index any files which can be accessed from the administering server.
Metadata myths are not true	<u>Myth:</u> Metadata is costly, whether implemented manually or automatically. <u>Reality:</u> The most dynamic automated metadata systems cost a fraction of other methods at the enterprise level. Furthermore, automated metadata does not replace manual implementation. Rather it augments manual implementation for cost-effectiveness and maximum search and retrieval accuracy. <u>Myth:</u> Metadata cannot ensure accuracy unless a large number of meta tags are embedded into each document. <u>Reality:</u> On the contrary, metadata has proven to be the most accurate of all document management and retrieval methods. Metadata allows a user to search a document collection by an innumerable number of field names, or meta tags, to allow for maximum relevancy. The most efficient document collections are managed using automated and manual methods.
Manipulate metadata easily	Metadata can be added and removed from documents. In addition, you can add unlimited data to meta tags. Meta tags can be customized for organizations, and carry unlimited potential for search and retrieval accuracy.
Manage maturing document collections	Statistics support the estimate that corporate document collections double every three months. Moreover, documents themselves mature. Metadata maintains growing document collections and maturing documents.

Table 3-1 – Benefits of Using Metadata

Ease the transition from paper to e-files	Many organizations are using specialized scanning techniques to convert paper to electronic files. Metadata eases this transition by search-enabling scanned documents, including text, PDF, and image file formats.
Language-independence	Other knowledge management systems struggle to establish multiple language support for global organizations. Metadata, using universal bibliographic standards, features tokens. Tokens overcome archive and retrieval inconsistencies caused by semantics and multiple languages. A token represents the idea, not the group of words that describe it. It doesn't matter if you drive a car or an automobile, or if you are transported in a coach. Tokens alleviate this concern, while other methods of knowledge management fail to meet user demands.
No implementation time or training required	Because metadata integrates readily with existing archival methods, there is no training required for organization members, nor is there an implementation period, or "transition" time, when an organization begins using metadata.

Once files are in electronic format, document management can further increase and enhance productivity. Options range from a simple file storage system with limited revision tracking, to a system that securely controls viewing, editing, and distribution of all information. The use of keywords and computing power to search valuable metadata attributes in an electronic system offers enormous time and efficiency benefits over the use of manual labor to search boxes, filing cabinets, or stacks of paper. Once found, any document in the system can be instantly delivered to the user's computer workstation without fear of its inadvertent loss or destruction.

3.4 Reproduction

The benefits of converting hard copy documents to electronic formats will not only reduce document storage and retrieval costs, but will also reduce document reproduction costs. Once documents are converted to electronic format and indexed for search and retrieval, document reproduction can be reduced to simple printing or plotting, reducing risks and increasing employee productivity.

Electronic documents can be infinitely copied, reproduced, and transmitted without any risk of damage to or alteration of the original data. Integrating a viewing and printing application into individual workstations is a simple, inexpensive way to link scanning with the ongoing build of a total document conversion solution. A small investment in a capable viewing and printing software package offers immediate benefits with little capital outlay and minimal training time. The right viewer can help increase access to information, speed time to completion, streamline workflow, and accelerate review cycles and change requests.

3.5 Additional Applications of Document and Content Capture

3.5.1 Support for Future Systems

Supports Electronic Document Management

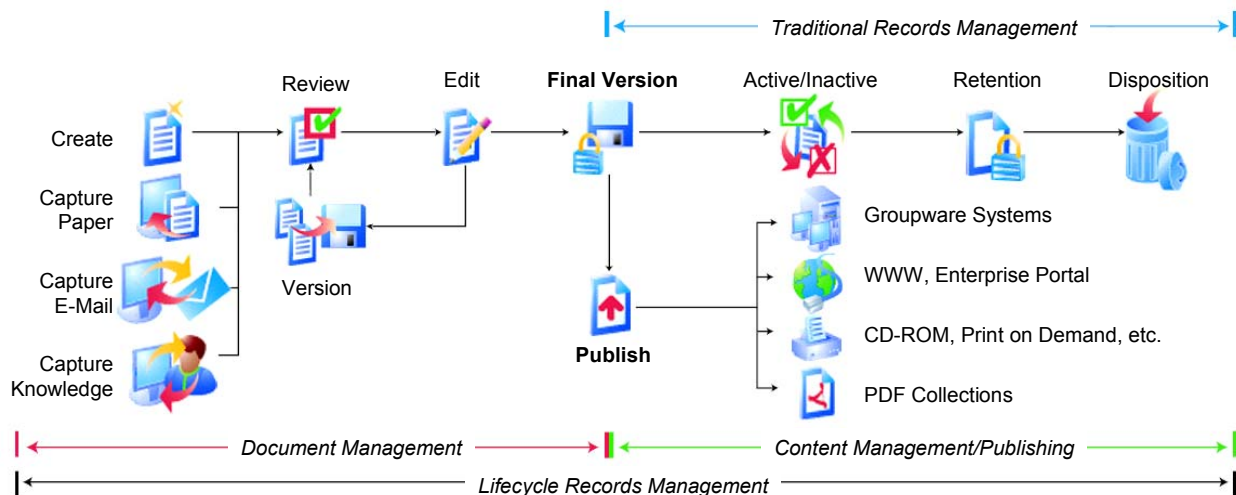
Converting legacy information to electronic formats and capturing metadata about that information leads naturally into building Electronic Document Management System (EDMS) content. Electronic document archiving, to be of value, has to be structured so that the user can efficiently retrieve the required document for the query or task being undertaken. Records and

document management both provide this support. An archive may require management by either or both disciplines dependent on the archive content. An EDMS provides general controls over storing and accessing electronic documents – a function well supported by a planned, methodical capture and conversion process.

Supports Records Management

A record, whether in paper or electronic form, is a statement of a business activity that must be preserved without change and that holds a description of the contents and the status of that record. Anyone in the enterprise, from the highest ranked official to the lowest ranked clerical worker, can create an information object satisfying the definition of an enterprise record. Records managers, in addition to being end users, manage records once they have been created. The management entails maintaining the records in recordkeeping systems that capture, manage, and provide access to records through time; using the records in the sense of making them available for active use in support of business operations; and retiring the records, when no longer needed for active use, to temporary or permanent storage.

Figure 3-2 – Lifecycle Records Management



A Records Management System (RMS) such as the Army Records Information Management System (ARIMS) provides the control and management functions to standardize file plans, record metadata and disposition schedules for an organization, and provide comprehensive auditing and control over the access and actions on records. Correctly deployed and integrated into the workflow, ARIMS is expected to provide for the management of both electronic and paper records.

ARIMS is intended to replace the Modern Army Recordkeeping System (MARKS). The new system will include a Web-based set of applications and tools, new disposition standards, and Army consolidated records holding areas (RHAs) to help in the management of hard-copy and electronic records. In addition, ARIMS Web-based components will include the Army Electronic Archives (AEA), a large-scale facility that provides for long-term, secure storage of electronic

records such as e-mail. ARIMS, programmed and Army funded through Fiscal Year 2009, will be available without charge to the Corps as part of the Army Records Management Program.

3.5.2 Disposition Program

A records disposition program provides for the effective and efficient management of records no longer needed in office space to conduct current business. It has three main objectives: (1) the prompt disposal of temporary records whose authorized retention periods have expired, (2) the timely and systematic transfer to economical storage of records no longer needed in office space but not yet eligible for final disposition, and (3) the identification and transfer of permanent records to legacy archives for preservation and for reference and research use.

Electronic records need not be kept permanently or even for long periods. All records, regardless of media, fall into one of two categories for disposition purposes:

- Temporary records. These should be destroyed, or in rare instances donated, after a fixed period of time or after occurrence of a specified event. The time may range from a few months to many years. Most Federal records are temporary.
- Permanent records. These are sufficiently valuable for historical or other purposes to warrant continued preservation by the Federal Government. Relatively few Federal records are permanent, although the exact percentage differs from agency to agency; the Corps has a relatively high percentage of permanent records.

In carrying out a records disposition program, electronic document retention periods for temporary records and suitable transfer dates for permanent records should be established through metadata. There is also a need to determine where the records should be kept as they await disposal or transfer to permanent archives: on the office server, in a separate storage facility, or in a commercial record center.

3.5.3 Knowledge Management

Knowledge management (KM) combines both technology and strategy. KM makes use of the technologies, tools, and methods used to capture, manage, store, preserve, and deliver content across an enterprise to build a strategy for managing an organization's information. At their most basic level, KM tools allow the management of an organization's unstructured information, wherever that information exists.

At the core of any effective business or enterprise application is the strategic management of content. Handled strategically, a well-planned knowledge management strategy can bind together the user-facing layer of information gathering with the "back-office" layer into an effective application.

A KM strategy needs to address many issues, including:

- How much content is there? How fast will it grow?
- How many types and classifications of information content are there?
- Who manages which content?

- Who owns the content?
- What is the content's lifecycle? What must be saved, in what form, and for how long?
- How does content get re-used and re-purposed? What standards should be used to ensure exchangeability?
- How do you optimize your strategies across processes so that there is consistency of approach and sharing of information?
- How do information and document repositories interoperate with applications to provide content?
- How do you document (with certainty) the context in which document transmission takes place?
- What is the electronic content migration strategy as technologies become obsolete?

A comprehensive knowledge management strategy considers all the following:

- *Capture*: Data capture, imaging, scanning and hybrid systems, color scanning, forms processing, handwriting recognition, character/voice recognition
- *Management*: Business process management and outsourcing, workflow, portals, search engines, content categorization, document management, digital asset management
- *Storage*: E-mail security and archiving, data warehousing, storage systems (Network Attached Storage (NAS), Storage Area Networks (SAN), Write Once/Read Many (WORM), Redundant Array of Inexpensive Disks (RAID), Magneto-Optical (MO), etc.), compression
- *Preservation*: Records management and archiving, film-based imaging, digital preservation, disaster recovery and planning
- *Delivery*: Digital printing and publishing, print systems and utilities, encryption, authenticity, digital signatures, electronic bill presentment, wireless/Bluetooth, Web content management

3.5.4 Document Workflow Management

A document workflow system is designed to automate a business process for the creation, updating, review, approval, and distribution of documents. Workflow can also provide a much-needed structure to an organization's document collaboration.

Document workflow systems generally support two workflow types: dynamic and predefined.

- Dynamic workflow is well-suited for small to mid-size teams collaborating on documents. It allows users to create and manage their workflow very quickly (and easily) without having to know the entire process ahead of time. This, in turn, provides maximum flexibility.

- Predefined workflow provides a powerful, high-volume mechanism to automate and prioritize document management processes. It can support processes that include either many tasks and/or a large number of participants. This kind of workflow supports a series of logical steps, each with a specific task and assigned participants. In addition, predefined workflow offers the advantage that business users can initiate a document process without knowing all the details. Users assigned to workflow tasks, for example, are presented with a concise request regarding the task or activity requested of them. Knowledge about previous or future steps in the workflow is not required. And, with this type of workflow system, subject matter experts can be brought into the appropriate stages during the document lifecycle to review and/or approve different document types.

A document workflow system serves to support some type of workflow template. This workflow template allows your organization to predefine and automate the document review and approval process you want to use. Workflows can be created for different document types as well as for organizational processes themselves. You can, for example, create a workflow that defines how press releases are created, reviewed, approved, and published. Another template can define how and at what stage a document is to be approved by your legal department before being routed to the corporate portal.

A document workflow system enables business users to easily create and modify workflows according to their specific business needs and processes. Pre-defined workflows further enable organizations to capture their business process, ensuring that documents are created, reviewed, approved, and published according to company policies and procedures. An effective document workflow system will require no programming and is ideally managed through a browser.

4 USACE Current Situation

4.1 Site Visit Data

Part of this study is based on information gathered from visits made at three individual Corp District sites. The goal of each visit was to gather information on what types of documents were to be converted and to get a better idea of the quantity of each type of document that would be included in the conversion process.

The following sites were visited:

- Ft. Worth District
- Mobile District
- Seattle District

Before each visit, each site was given an ADCS Site Interview Document (See Appendix G for a draft example). The site interview document was to be used as tool for communicating Corps of Engineers' requirements for drawing conversion in support of their Military Construction program. Questions and general topics were gathered to encourage communication during the site interviews. They are as follows:

- Are there any initiatives underway to convert documents and drawings?
- Probable future need for documents and document types
- Specify non-digital or digital formats currently in use (paper, Mylar®, aperture cards, microfiche, CALS Type 1 and/or other vector, TIFF and/or other raster, outdated CAD file formats, etc.).
- What disciplines do the drawings represent and in what formats are they currently stored?
- Describe the condition of drawings (deteriorating, negatives unusable, etc).
- How many documents/drawings would be considered for conversion?
- List file formats required for converted documents and drawings.
- State accuracy requirements of converted document and drawings.
- Specify adherence to CAD standards used during conversion.
- Specify anticipated needs for documents and drawings outside of local district.

Much of the information collected at each site was centered on the various documents types and quantities of each. Some sites reported better than others. Steps were taken to view actual file rooms / storage areas and gather at least a visual estimate of how many documents each site had in inventory.

4.1.1 Ft. Worth District Notes

Table 4-1 – Ft. Worth Document Types and Quantities

Document Type	# of Pages	Format Size	File Format	Indexing	Location
Aerial Photo Negatives in boxes	5,500	9×9	Negative Reels	Stacked in boxes	3rd Floor File Room
Drawings (large, photos, maps) #1	65,000	Varies	Paper	Stacked in drawers	Survey Room
Drawings (large, photos, maps) #2	312,000	Varies	Paper	Stacked in drawers	Survey Room – File cabinets drawers (52)
Drawings Mixed	672,000	Varies	Paper	Stacked in drawers	Survey Room – Large stacks
Flood Reports	138,600	Bound, 11×17	Paper	Shelved	3rd Floor File Room
FPMS Docs 1	75,000	Varies	Paper	File cabinet	3rd Floor File Room
FPMS Docs 2	84,000	Varies	Paper	File cabinet	3rd Floor File Room
FUDS	192,000	Varies	Paper	File cabinet	3rd Floor File Room
Gage Books	336,000	Mostly 8½×11	Books/ Paper	Stacked in a cabinet	3rd Floor File Room
H&H Studies	22,500	8½×11	Paper	File cabinet	3rd Floor File Room
H&H Study Maps	15,000	C and E Size	Paper	Stacked in drawers	3rd Floor File Room
Misc. Maps	3,000	Varies	Paper	Stacked in drawers	3rd Floor File Room
Quad Maps (double stacked)	17,500	Varies	Paper	Stacked in drawers	3rd Floor File Room
Reports Archives Survey GPS Control	84,000	8½×11	Paper	File cabinet	3rd Floor File Room
Reservoir Control Archive	126,000	8½×11	Paper	File cabinet	3rd Floor File Room
Survey Field Books	240,000	Bound, 8½×11	Books/ Paper	Shelved	3rd Floor File Room
Survey Hydro Maps, Mosaics	15,000	Various	Paper	None	3rd Floor File Room
Survey Notes Boxes	178,200	8½×11	Paper	Stacked in boxes	Survey Room – Survey Note Boxes (99)
Topo Maps	7,500	E Size	Paper	Stacked in drawers	3rd Floor File Room

There are approximately 2.6 million pages located on the 3rd floor file room. The 4th floor file room was not inventoried in detail but visual estimates indicate it has approximately twice as many pages. This total of approximately 7.8 million pages does not include documents stored offsite.

Departments Interviewed:

Records Management – Andrew Goss (maintains inactive records)

- There are approximately 768 boxes on the 2nd floor of the building and 6905 boxes located in the Federal Records Center.
- They have a new indexing system written in Visual FoxPro.
- Users will soon be given online access.

- Presently, the new indexing system has only 138 records available. More records need to be indexed.

ENC (Electrical and Civil Engineering)

- This group contracts out all survey and mapping work to subcontractors. The contractors then deliver all information in electronic format. Information is stored with project information in file cabinets.
- File room consists of 120 drawers of project files. No as-builts. (They say that as-builts are not worth keeping) 50 drawers of old site maps for installation. Electrical has 25 drawers. There are 44 single drawers of (as-built) microfiche. Approximately 1600 sheets per drawer.
- AE designed drawings are put on server. For AE contracts, CD goes to contracts for the official record.

Hydrology and Hydraulics

- This area contains 3 shelf cabinets with 400×16 sheets of drawings. Should be copies in Construction.
- Have approximately 100 drawers of flood maps.

4.1.2 Mobile District Notes

Table 4-2 – Mobile Document Types and Quantities

Document Type	# of Docs	Format Size	File Format	Indexing	Location
GIS – Maps	Quantity Unknown	Various	Various	CAD and GIS systems	On-site
Hydropower plant drawings A	6,000	Various	6000 drawings converted to tiff images	No indexing	On-site
Hydropower plant drawings B	114,000	Various	All drawings in paper format	No indexing	Hard Copies stored in basement. Tiff images are on CDs.
Planning	Quantity Unknown	Various	Various	No Indexing	On-site
Real Estate – Microfilm	400,000	Various	Microfilm	No indexing	On-site and duplicates stored off-site
Regulatory documents – Microfilm	616,000	8 ½ x 11 images	308 rolls of microfilm – 2k per roll	No indexing	On-site
Regulatory documents – Permits	Quantity Unknown	Various	Tiff images	Have a scanning system like CEERIS	On-site

In the departments visited, there are approximately 1.2 million pages. Not every department was interviewed and time did not permit the counting of the file room(s). This total of approximately 1.2 million pages does not include documents stored offsite.

Departments Interviewed:*Information Management –Audrey Imsand*

- Vault Room contains 300 drawers of paper documents. It is estimated that there are approximately 400 sheets per drawer. This equals a total of 120,000 documents located in the vault room that need to be scanned.

Real Estate Department –Bud Lee

- The Real Estate department operates with drawings in Microstation format for which TIFF images are scanned @ 600 dpi. No quantities were collected.
- There are 400,000 survey maps stored on microfilm that they scan as needed.

Design – Ed McLauren

- They have a tape archive which can be searched to tell the user which box to go look in to find the drawing being searched. In other words, each box contains a stack of drawings.
- No quantities were collected.

Hydro Powerplants –Les Harper

- A large portion of the drawings at all powerhouses have been scanned and converted into a TIFF format. Approximately 6000 drawings (in TIFF format) exist and are stored on CD. No mention of any indexed system or way of finding a particular drawing from any CD.
- There are 114,000 flat file drawings in records management in the basement of the Mobile District building. There exists a potential data loss hazard here because the 1st floor is susceptible to floods and there is a risk here that drawings could be damaged.

Regulatory Branch –Jim McEnry

- The Regulatory Branch completed a scanning project that automated their records management process. This was initiated to meet legal requirements for admissibility and trustworthiness. The system improved timeliness in record and document access, improved quality of responsiveness, increased productivity, and reduced the cost of operations. The system cost was estimated at \$9,252; much more attractive than the CEERIS cost estimate of \$58,400.
- No estimates on quantities of documents scanned were collected.

4.1.3 Seattle District Notes

Table 4-3 – Seattle Document Types and Quantities

Document Type	# of Pages	Format Size	File Format	Indexing	Location
Aerial Photo Negatives	250,000	9x9 inch negative	Reel Film Negative	Separate database	Map Records
Aerial Photo Prints	27,000	9x9 inch print	Photographic Print	Boxes are indexed in database	Off site boxes
Drawings – Aperture Cards	90,000	Aperture card	Aperture card	File cabinet	Project Office
Drawings – Maps, charts, etc.	120,000	Aperture card	Aperture card	File cabinet	Project Office
Drawings – Projects	37,500	E Size	Mylar and TIFF	Indexed by file cabinet and drawer – database not usable	Project Office
Operations -Survey Books	100,000	8½ × 11	Paper	Manual	File cabinet
Real Estate Legal Descriptions	5,000	8½ × 14	MS Word or TIFF	Indexed in REMIS	Stored on LAN
Real Estate Legal Docs	50,000	8½ × 14	Paper	Not scanned or indexed	Real Estate file folders
Real Estate Property Maps – Retired	2,800	E Size	½ in Mylar Sepia, ½ in MicroStation	Manual	Hanging in R.E. file folders

There are approximately 680,000 pages stored by the departments surveyed. This total includes only 27,000 of the pages stored offsite.

Departments Interviewed:

Real Estate – Steve Mortensen

- 1100 retired drawings, 1700 active drawings (E-size).
- Workers access about 100 sheets/day and change 10 drawings/month (33% of the drawings are in CADD systems).
- When a drawing modification is needed and that particular drawing is not in CADD, RE contracts out the vectorization to a firm in Portland, Oregon and gets in return a Microstation file. Then modifications are made to the Microstation file.
- Scanned old legal documents and some MS Word docs / stored on disc (on LAN). Indexed in REMIS by legal description (only comments field is indexed). 1526 documents exist past 1984; 933 documents exist prior 1984.
- Unknown quantity of historical files

Aerial Photography

- 230,000 images (film on rolls).
- 33 boxes of contact prints do not exist as negatives and are currently being scanned at 300 dpi.
- Aerial Design – 250,000 in quantity. Access 5 cans/day, growth rate 5 cans/year.
- Surveillance – 250,000 in quantity. Access 5 cans/day, growth rate 15–18 cans/year.
- 30 or 40 boxes of aerial photos are starting to deteriorate/fade. Starting to scan at 900 dpi. Indexed by project.

Operations (Navigation Drawings) – Lonnie Reid-Pell

- Navigation drawings stored in map file records; now creating in Microstation. There exists a need to access via Web interface. Scanning engineer: Joyce.

Hydro Power

- 99% scanned. Most on server, indexed in database organized by file cabinet & drawer from original hard copy. Uses Access as the database.
- Files on server in same structure as hard copy.
- There are unspecified problems with the database.

Dams

- Libby Dam – 10,000 – 90% scanned
- Chief Joe Dam – 10,000 – 99% scanned
- Howard Hansett Dam – 2,500 – 0% scanned
- Mud Mountain Dam – 1,000 – 0% scanned
- Albany Falls Dam – 10,000 – 10% scanned
- Lake Washington – 2,000 – 0% scanned
- Wynoochie – 2,000 – 0% scanned
- Drawings on Aperture Cards – 90,000 – have scanned 30,000 – 33% scanned

4.2 Site Visit Observations

The specific site visit data, along with anecdotal evidence gleaned during the visits, led to several wider conclusions about the Corps' current document management situation. The Corps' clear success in document management with Electronic Bid Sets (EBS) can be directly attributed to two factors: (1) EBS was implemented uniformly throughout the Corps, and (2) EBS involves

mandatory, software-supported metadata collection. However, apart from EBS, as a general matter a lack of high-level vision for document management or metadata collection results in inefficiencies, lost time, and reduced productivity throughout the Corps.

Mountains of Paper

The following general observations were made:

- Lack of consistency in indexing methods for documents (names, projects, properties, etc) lengthens document searches, delays information retrieval, and imposes costs on Corps processes.
- Indexing methods are generally not consistent between districts.
- No district-wide search capability exists; locating a document is limited to a division, a department, or even a section within a department, and there is generally a need to locate a specific individual to perform any search (with limited exceptions).
- Though increasingly electronic workflow is not total, increasingly lower priority is being assigned to a new paper created.
- Paper continues to be generated without common indexing or electronic capture, exacerbating the existing problem. The amount of paper-only data is as large as might be expected given the Corps' long history.

The following storage scenarios were either directly observed or related anecdotally during the site visits:

1. Some departments store and keep up with their own documents for work in progress / work recently completed.
2. Some departments keep up with documents throughout the documents' lifecycle. For example, Real Estate legal descriptions and legal property maps are kept current.
3. In other cases, departments only keep up with documents while the project is ongoing, then store the documents elsewhere external to the department.
4. Some departments or divisions have a records storage area within which all documents for department/division are indexed and kept. These storage areas only apply to a single department or division and are not integrated across district.
5. Some district-wide document storage areas (file rooms with flat files and file cabinets) are maintained.
6. All districts have access to offsite storage facilities (local record building areas or extended file storage areas) to house box quantities only.
7. Aerial photograph images (negatives and positives) are generally stored as 9"× 9" negatives on continuous reels. In most cases, there are no duplicates.

These storage scenarios tended to lead to corresponding, highly individualized search and retrieval schemes:

1. Many documents are either stored in file cabinets (indexed by current project) or stored on some workers' desk. Indexing is almost entirely ad hoc and subject to the memories of the personnel involved.
2. In some cases, a fairly well thought-out indexing scheme for department-wide data catalogs documents that will be used frequently and/or on an ongoing basis. For example, Real Estate legal descriptions are indexed and kept in Lectrivers™, but this system is enforced by the Corps-wide Real Estate Management Information System (REMIS). Also, Design projects are commonly indexed by Project Number (PN), but some districts do a more consistent and/or thorough job than others of capturing location and type of project in the PN than others.
3. Departments are especially likely to keep up with documents only until project conclusion if the final office of record for the document is external to the district. Anecdotal information indicates that the office of record loses documents uncomfortably often, and that that office then requests additional copies of the lost documents from the originating department in the hopes that originals were kept (contrary to policy).
4. The department or division records storage area usually provides a more predictable method of finding and retrieving documents, but indexing is based on the particular needs of the departments involved and tends not to address the needs of the entire district.
5. The search capabilities of these storage areas are mixed. In some instances, documents appear to be organized predictably (though external indexes are used to find documents could stand significant improvement). In other instances, large file rooms are largely indexed, and locating documents relies on the personal memory of individuals.
6. Storage facilities index boxes only, making the ease of finding any given document uncertain. Usually, locating and retrieving a number of boxes in which the sought document might have been stored is only the first step; the boxes must then be manually searched.
7. Photography reels are filed by the contract number that authorized the flight (and not by geography). Assuming the reel can be handled at all, it must be searched manually to locate the appropriate image. In addition, the process by which aerial photograph reels are read involves a significant chance of damaging the original negative and introducing flaws into or even obliterating irreplaceable (often historically and environmentally significant) imagery.

Uncertain States of Preservation

In addition to the general understanding that documents deteriorate over time, the following observations were made:

- Blueline prints are generally created by an ammonia process that is very sensitive to sunlight (the blue lines fade to white).

- Over time, blueprint prints, linen drawings, and vellum drawings tend to accumulate water marks, food and drink stains, and pencil, pen, and eraser marks.
- Paper warps easily and loses dimensional fidelity, particularly if pinned to a wall or exposed to uneven moisture.
- Mylar[®] plots (especially those produced by electrostatic devices) can be accidentally smeared or even wiped clean of markings when handled improperly.
- Aerial photograph negatives are suffering deterioration to near-destruction due to suboptimal storage conditions and handling.

At best, the Corps' extensive library of aging documents is becoming progressively harder to preserve and use over time. In practice, many of these documents are in danger of complete loss within only a few more years.

Anonymous CD-ROMs

Site visits and anecdotal information revealed the following issues with documents that have already been scanned, whether within the Corps or by outside vendors.

- Stacks of CDs reside in individual worker's cubes, but their contents are not indexed and cannot be ascertained short of depending on individuals' memories or manually searching the files on the CDs.
- As departments scan documents and build up individual collections or stacks of CDs, no steps are taken to share that information. Users cannot access information outside of their department.
- File names on the CDs tend to be both nondescriptive and so generic that they are easily overwritten; *e.g.*, 0000001.tif etc.

Outdated, Outmoded Electronic Formats

Data still exists in the Corps on all of the following physical media:

- Legacy magnetic tape (VAX and UNIX)
- Legacy optical media
- 8" and 5¼" floppy disks

While the media artifacts themselves may be perfectly sound and may still contain data, in many cases the machines that can read them are gone or no longer function. In addition, compatibility problems, such as incompatibility between old versions of AutoCAD[®], increase the effort and time, and therefore cost, it takes to access historical data workers need to do their jobs.

Metadata Issues

Site visits and anecdotal information tend to indicate that, where collected, document metadata is inconsistent at best. In general, Corps processes tend not to collect metadata reliably often. When

metadata collection is required, divisions and departments tend to act alone in originating idiosyncratic metadata requirements, which they then implement in systems that see only limited use. This metadata may appear in manual indexes for hard-copy documents or in electronic indexes (such as Microsoft® Excel® spreadsheets) for documents in electronic formats.

One exception to this rule is the metadata collected for the indexing of an EBS. An important factor in the success of the EBS scheme, mandated for use across the Corps, is its use of third-party software that requires a consistent set of metadata in order to generate an EBS.

Reusing Information

In general, information is reused in the Corps on a regular basis, and some Corps projects would benefit substantially from having easy and fast access to the Corps' archived data. Site visits and anecdotal information revealed the following among Corps workers' attempts to reuse electronic information the Corps had previously converted or collected:

- Where there are scanned documents, there tends to be no association between the scanned electronic documents (whether scanned for EBS or for other reasons) and any source data used to generate those documents in usable text or vector (Word, CADD, etc) format.
- Raster formats are immensely difficult to reuse or to extract data from.
- Suboptimal scans, whether due to equipment or procedure, cause data loss. For example, aerial photographs scanned at 300 dpi are useless for their original purpose, because the pixel size of an aerial photograph is around 12 microns; thousands of dpi.

4.3 Conversion Challenges

The observations made and information related during the 2002 site visits present the following items as challenges the Corps must overcome when implementing ADCS.

4.3.1 Paper

Paper production within Corps is continuous. Each day new records are being stored away in file rooms, file cabinets, or external storage facilities. Each year the Corps stores millions of documents of various sizes. To improve budget performance, the Corps has identified a need to decrease, not increase, the costs associated with paper storage. Still, there are requirements for documents to be kept in storage and readily accessible. Every year the Corps pays for storage space to house these documents. The Corps has determined that reducing or eliminating paper storage will significantly reduce cost of floor space within the organization. The Corps' Headquarters site estimated in 1999 that it could save about 1 million dollars per year by eliminating the need for hard copy document storage.

The Corps has also found that retrieving documents for review or reproduction can be a tedious and time-consuming task. Documents can be stored at various locations and retrieval can be difficult to determine. Anecdotal evidence indicates that workers spend many hours searching through boxes of stored documents and that needed documents are not always accessible. In some instances, needed information is never found and must be laboriously reproduced if that is even possible. These accessibility issues increase the cost of doing business within the Corps.

4.3.2 Deterioration

The materials of which the Corps' document collections are composed – paper, vellum, linen, leather, and adhesives used in bookbinding – are susceptible to two main forms of deterioration. One is biological deterioration caused by insect attack and/or fungal growth, and the other form of deterioration is caused by adverse environmental conditions such as extremes of dampness or wide fluctuations in relative humidity associated with large variations in day and night temperatures, light and atmospheric pollutants. These two forms of deterioration are interconnected because humid conditions favor the growth of fungi and accumulations of dust and dirt will attract insects.

Biological Organisms

Condensation or moisture due to high humidity encourages biological growths such as molds or fungi. Also, insects and even rodents tend to inhabit infrequently-used areas such as document archives. Biological agents attack paper and other organic materials when temperature and humidity are uncontrolled. Mold growths digest the material on which they grow, resulting in staining, deterioration, and rapid loss of strength in organic materials. Mold growth occurs most readily on items made of organic materials that are tightly packed, because this forms a thin, stagnant pocket of moist air which favors mold growth.

In addition to high temperature and humidity, other conditions also favor the growth and proliferation of insects. The following are some of these conditions:

- Accumulations of dirt and dust from poor or careless housekeeping practices
- Introduction of foodstuff into working areas
- Open windows, air vents or poorly sealed windows and doors
- Unattended roof leaks and cracks in a deteriorated building
- Poor ventilation

Rodents and insects are the worst enemies of paper and other organic materials that are cellulose in nature. Paper, especially books, and other materials may contain proteins and carbohydrates in the form of sizing, paste or starches, and other organic substances attractive to vermin. The nature and extent of the damage depend not only on the creatures and materials involved, but also on how promptly the infestation is discovered and controlled. Damage may vary from a few holes to complete destruction.

The most common types of insects that attack paper objects are:

- Termites live in wood and under the ground under conditions in which humidity within the colony is maintained at a high level. In paper archives, they can produce deep, crater-shaped holes, or deep, irregularly shaped erosions and bring about irreparable loss or damage.

- Silverfish are usually found in moist locations, that is, under stones and boards, cracks and crevices or in dark places where humidity is greater than 55%. They cause superficial damage to paper, eat away glue, paste, etc., and attack photographic plates and gelatin.
- Cockroaches hide in warm, damp, and dark places like bathrooms, kitchen floors, near water pipes, in crevices, cabinets, and cupboards. On paper, they can cause superficial erosion of irregular outline; a blackish “comma” shape mark on paper is a positive indication cockroaches are present.
- Booklice cause tiny superficial erosions of irregular outline to paper, leather, gelatin of photographic plates, watercolors, parchment, and glue and gum of bookbinding.
- Case-bearing clothes moths thrive in undisturbed and unventilated areas and destroy bookbinding.
- Powder post beetles bore holes into books and other organic materials.
- Deathwatch beetles dig winding, circular tunnels that generally extend from a book’s edges to its center; the resulting powdery mixture that fills the tunnels is known as “frass.”
- Carpet beetles cause irregular perforations and sometimes surface tunnels containing powdery remnants on books and other paper organic materials.

Temperature and Relative Humidity

Temperature and relative humidity are interdependent as damaging factors. Hygroscopic materials (those that normally contain moisture) are very sensitive to both humidity and temperature. Those hygroscopic materials of organic origin and of fibrous or cellular structure, such as paper, parchment, papyrus, leather and the adhesives used in bookbinding, deteriorate rapidly with temperature and relative humidity changes.

The greatest danger from high relative humidity is the tendency for molds to grow on any material that providing nutriment, such as glue, leather, or paper. High humidity also hastens acid deterioration. Mold growth is a warning that the atmospheric relative humidity is above the safety limit for the paper in the room. When conditions are favorable to mold growth, a gray dusty bloom is observed, at first on darker bindings, but soon becoming fluffy with a tendency to be organized in circular patches. This sort of damage can make a document unreadable.

The greatest danger from low relative humidity is over-drying, which causes hygroscopic materials to weaken and become brittle. In extreme cases, handling over-dried hygroscopic materials can cause them to disintegrate, taking with them the printed information they stored.

Regular changes in temperature and relative humidity (cycling) can also lead to weakening of paper and related materials, as a result of internal stresses set up in the material in response to these changes. While there are no firm data to indicate how serious this effect may be, preservation scientists believe the damage can be minimized if changes in temperature and relative humidity can be held to less than ± 10 degrees and $\pm 15\%$ over the lifetime of the archive. These conditions are difficult to establish without expensive climate control systems, and many of the Corps’ document storage areas probably do not meet this standard. The passage of time is rendering more and more of the Corps’ collected knowledge fragile and in danger of loss.

Light

Apart from other causes earlier discussed, mere light can be regarded as an independent and prime cause of archive deterioration. Materials especially subject to damage by light are pigments and dyestuff, including inks; paper and other cellulose materials; and various other organic materials.

Pigments and dyes fade when exposed to light; this is very noticeable in watercolor paint, but is also quite pronounced in printed output. Unfortunately, colors fade selectively, some disappearing while other remain relatively unchanged, which means that the color data on a multicolor drawing, for example, can be partly lost.

Rapid and serious deterioration of the paper itself is caused by the cellulose oxidation brought about by ultraviolet rays present in both sunlight and fluorescent light. There are two effects of light on paper that result in its ultimate embrittlement and deterioration. First, light has a bleaching action that causes some whitening of paper and fading of colored papers and certain inks. Second, it causes any lignin, which may be present in the paper, to react with other compounds and turn yellow or brown. (It is this reaction that results in newspapers' turning yellow with age.) Certain invisible changes also occur at the same time when these visible effects of light are taking place. Paper fibers fragment into smaller and smaller units until they are so short they can no longer hold the paper together; eventually, light-damaged paper will also disintegrate into useless powder. Unfortunately, the reactions initiated by light continue after the source of the damage has been removed.

Atmospheric Pollutants

Finally, materials of organic origin such as leather, parchment, and paper products tend to be soiled and stained by solid particles of carbon, tarry matters, and other solid contaminants present in the atmosphere. The worst contaminants for this group of materials are the sulfurous and sulfuric acids that result from the combustion of fuels and from other industrial processes. The effects are most severe in cellulose materials such as paper and leather. There is a close correlation between the loss of strength of paper and its acidity resulting from sulfuric acid contamination. Dust and dirt particles in the air not only carry with them the adsorbed pollutants mentioned above but may exert an abrasive action on books and paper. Expensive filters can ameliorate this problem, but it is not known how many Corps document storage facilities circulate only filtered air.

4.3.3 Volume

A big obstacle to managing information in the Corps is the sheer volume of information the Corps has available. The time or duration of any document conversion project increases (usually linearly) with volume. Increasing manpower to decrease time has no effect on overall cost, assuming two workers cost exactly twice as much as one worker.

Volume often hinders progress in any conversion process because the funding needed to complete a large-volume project is prohibitively high. In May of 1999 the Corps of Engineers started its Corps of Engineers Electronic Document Management System (CEEDMS) initiative. CEEDMS was an effort to institute an enterprise-wide electronic document management system. The initiative was not completed, but the Corps did estimate the total number of documents at

each site that needed conversion: it was estimated that over 464 million pages would be included in the document conversion process. The table below shows the estimated total number of pages per field site.

Table 4-4 – 1999 Document Estimate per Field Site

Organization	8½ × 11 & 8½ × 14 TIFF/PDF Pages	E & C Sized TIFF/CALS Pages
LRH (Huntington) Test	7,500	2,500
MDC	349,440	14,560
PODHQ	510,720	21,280
LRDHQ2	672,000	28,000
NWDHQ2	1,182,720	49,280
NADHQ	1,344,000	56,000
SADHQ	1,384,320	57,680
LRDHQ	1,585,920	66,080
SWDHQ	1,626,240	67,760
NWDHQ	1,680,000	70,000
SPDHQ	1,800,960	75,040
SAC	1,948,800	81,200
WRC	1,962,240	81,760
MVDHQ	2,136,960	89,040
CPW	2,150,400	89,600
UFC	2,701,440	112,560
HECSA	2,714,880	113,120
LRC	2,903,040	120,960
LRB	3,642,240	151,760
POF	3,830,400	159,600
SPN	3,884,160	161,840
POJ	4,072,320	169,680
SPA	4,287,360	178,640
CRREL	4,368,000	182,000
TAC	4,529,280	188,720
CERL	4,583,040	190,960
TEC	4,583,040	190,960
SWG	4,932,480	205,520
NAU	5,040,000	210,000
SAW	5,187,840	216,160
POH	5,443,200	226,800
POA	5,725,440	238,560
NAO	5,940,480	247,520
LRE	6,773,760	282,240
NAP	7,123,200	296,800
NAE	7,674,240	319,760
HNC	7,674,240	319,760
MVM	7,969,920	332,080
NWW	8,077,440	336,560
NAN	8,937,600	372,400
MVP	9,488,640	395,360
SWL	10,147,200	422,800

Table 4-4 – 1999 Document Estimate per Field Site

Organization	8½ × 11 & 8½ × 14 TIFF/PDF Pages	E & C Sized TIFF/CALS Pages
MVS	10,268,160	427,840
NWS	10,402,560	433,440
SAJ	10,631,040	442,960
SPL	10,671,360	444,640
LRN	10,967,040	456,960
HQUSACE	3,874,129	336,881
LRP	11,760,000	490,000
NWK	11,854,080	493,920
MVR	11,894,400	495,600
SWT	12,458,880	519,120
LRH	12,633,600	526,400
SWF	13,130,880	547,120
SPK	13,762,560	573,440
SAS	14,434,560	601,440
LRL	15,294,720	637,280
NWP	16,141,440	672,560
WES	16,558,080	689,920
MVK	16,625,280	692,720
NAB	16,692,480	695,520
MVN	16,786,560	699,440
NOW	17,377,920	724,080
SAM	17,740,800	739,200
TOTALS	464,537,629	19,533,381

4.3.4 Standalone Storage

Search and Retrieval

One of the main goals of capturing and converting documents to electronic format is reducing costs associated with document and drawing retrieval. Anecdotal data adduced during site visits indicated Corps workers remain strongly interested in faster document searches and more reliable means of document retrieval. These goals can only be met by following a complete, detailed process when scanning documents and drawings, including a plan for digital permanence using electronic storage media. For best results, the Corps' scanning process should include a better mechanism than CD-ROM libraries for document storage.

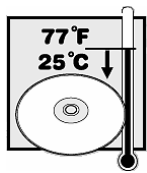
Though many departments have lots of information stored on CDs, most have no idea what document is on a particular CD, or how to quickly find a document related to any particular topic. The average CD-ROM library represents some progress in digital archiving, but without management of information or collection of metadata the work loses much of its value. In addition, the possibility of irreparable CD damage presents districts with a slim but always-present risk of irreversible information loss; it is not only the CD but the scanned information that is vulnerable because the original information is typically stored away in a warehouse and difficult to relocate. Without a distributed, network-accessible storage location for documents

and drawings, document retrieval across departments cannot be supported. Planning and consideration is needed to build a permanent repository of documents and drawings that can be accessed by Corps workers across departments.

Digital Permanence

Information contained only on a CD on a bookshelf is not the best permanent means of storing information. While nothing lasts forever, information stored on a file server can be backed up to streaming tape or other types of archival electronic media. CDs are expected to last a maximum of 70 to 100 years assuming average storage techniques. Heat is the worst culprit for this type of media. For best results in CD life expectancy, CD media must be stored in places where heat and humidity are moderate. Direct sunlight is another culprit to the life expectancy of CDs. Direct sunlight can damage CD media within minutes. The solvents in permanent ink are known to dissolve CD materials over time.

Although CDs can tolerate a fairly wide range of conditions without immediate damage, it is best to store CDs in cool and moderately dry environments. Slow chemical changes such as oxidation of metallic reflecting layers, dark fading of dyes, and deterioration in polymer substrates and coatings are temperature- and humidity-dependent processes. They will always proceed faster under warmer and more humid conditions. For Kodak writable CD products, for example, the manufacturer predicts the discs will deteriorate beyond usability in about 200 years, but only if stored at optimal, unvarying conditions.



In general, storage temperatures for CDs should not be any warmer than about 25°C (77°F) when archival permanence is desired. Cooler temperatures (down to about 10°C or 50°F) will help ensure the longest possible life expectancy. (It's neither necessary nor desirable to freeze CDs.) Relative humidity conditions should be in the range of 20% to 50% RH for best results.

Warmer and damper conditions mean lower life expectancy. Changing conditions, especially between extremes of temperature and humidity, can be dangerous to CDs. Although moderate temperature and humidity changes ordinarily do no harm, fast changes between very warm and wet conditions to cooler and very dry conditions might produce warping and distortion. The recommended maximum limit for temperature change is 15°C (27°F) per hour. For humidity, the recommended maximum RH change per hour is 10%. In practice, such steep gradients of temperature and RH are rare. CD cases or cabinets provide a modicum of defense against such cycling, ameliorating the rate of change in RH; the inside of a case or storage cabinet does not change as fast as the outside atmosphere.

At present, there are no national or international standards for long-term CD storage conditions published by either ANSI or the International Standards Organization (ISO). Until such standards are available, these recommendations must be considered provisional.

4.3.5 Legacy Electronic Formats

Enormous amounts of digital information are already lost forever. Digital history cannot be recreated by individuals, and organizations cannot recreate digital history if it is not archived or managed properly or if it resides in formats that cannot be accessed because the information is in outdated word processor formats, in old database formats, or on unreadable media. Many large

data-sets in governments and universities world-wide have been made obsolete by changing technologies (storage media as old as punch cards and 12" floppy disks) and will either be lost or subject to expensive "rescue" operations to save the information.

Information technologies cycle into obsolescence approximately every 18 months. This dynamic creates an unstable and unpredictable environment for the continuance of hardware and software over a long period of time and represents a greater challenge than the deterioration of the physical medium. Many technologies and devices disappear as the companies that provide them move on to new product lines, often without backwards compatibility and ability to handle older technologies, or the companies themselves disappear.

Document and media formats tend to multiply with time, and each one may carry its own hardware and software dependencies. Copying these formats from one storage device to another is simple. However, mere copying is not sufficient for preservation purposes: if the software for making sense of the copies (that is, for retrieving, displaying, or printing the files' contents) is not available, then the information will be, for all practical purposes, lost. Any document library must contend with this wide variety of digital formats. And many digital library collections originated not in digital form but from materials that were digitized for some purpose. Those digital resources which come to libraries from creators or other content providers will vary wildly in storage media, retrieval technologies, and data formats.

Libraries which take advantage of digital formats by cross-linking (including hyperlinking) or including multimedia contents will quickly discover the complexity of maintaining the integrity of links and dealing with dynamic documents, back-end script support, and embedded objects and programming that present their own format peculiarities and obsolescence cycle.

The challenge in preserving electronic information is not primarily a technological one (electronic information is relatively easy to transmit, copy, and store); it is a sociological one. The dynamism of the market for information technologies and products ensures the fundamental instability of hardware and software primarily because product obsolescence is often key to corporate survival in a competitive capitalist democracy. Product differentiation manifests itself at the very level of the document standard. Proprietary systems provide commercial enterprises with profitable products whereas static (*i.e.*, preservable) formats do not create the continual need for upgrading that software and hardware companies depend upon. This situation conspires against standards that would tend to create a stable nexus of hardware, software, and administration. Therefore, widely accepted standards that solve fundamental issues with respect to digital information will probably not emerge.

4.3.6 Metadata Collection

While state of the art hardware and software makes scanning easier, some adherence to a standard process should be mandatory while scanning. Many departments are giving only cursory thought to electronic document storage or naming conventions; the result is a cascade of filenames incomprehensible to most workers. Metadata is a collection of attributes stored in a database that describe or provide information about a document or drawing.

Metadata can be very simple or extremely detailed, but some information is needed to locate a particular document, information collection, data set, or service. Metadata is used to organize and maintain an organizations' investment in data. If content is inadequately described, it is

difficult to ascertain its value or use. The original investment is lost, since use of the data is limited. However, if the data has sufficient descriptive information, it is possible to determine its potential use, thus increasing its value, relevancy, and life span. Metadata helps in identifying data-integration problems that may exist in future systems. The proper definition and use of metadata can support future document management, records management, and knowledge management systems.

Many Corps Districts are currently scanning documents and information as needed. As observed during site visits, a large number of these scanned images are stored on CD, with little effort given to collecting metadata. The collection of metadata is often referred to as indexing. When indexing a document, a user will provide information about the document. This information is generally contained in a database and is used to help link documents to document attributes or characteristics that might be known by a user when searching for information. Since few departments are collecting metadata for their scanned images, there is no attribute information to search on when searching for a document.

Proper metadata collection is attributed to an organizations' careful research and a definition of all the metadata attributes that might describe the data. Metadata requirement studies are often carried out before any document conversion starts. Once defined, scanning software requirements can then be chosen to include metadata capture as apart of the document conversion process. Feature-rich scanning software can enable users to collect metadata before storing the document in a document repository.

Manual Indexing

Manual metadata collection can be time-consuming. Using the 1999 Corps estimates for the number of pages planned for conversion, the following facts were obtained.

(Assume that six fields of information need to be typed in for each record and it takes an average user five seconds per field to populate the field with the appropriate data. It therefore takes $(6 \times 5 =)$ 30 seconds for a user to enter all the metadata for that one record.)

- 1 workday = 8 hours \times 60 minutes \times 60 seconds = 28,000 seconds
- 1 work week = 40 hours \times 60 minutes \times 60 seconds = 144,000 seconds

Mobile (number of documents based on estimates from the USACE)

- $17,740,800 (8\frac{1}{2} \times 11 \text{ and } 8\frac{1}{2} \times 14 \text{ pages}) + 739,200 (\text{E \& C sized pages}) = 18,480,000$ (total # of documents at district location)
- $18,480,000 \times 30$ (seconds for each document to manually enter metadata) = 554,400,000 (total seconds for all documents to be manually indexed)
- $554,400,000$ (total seconds) \div 144,000 (seconds in a work week) = 3,850 weeks
- $3,850$ (total weeks) \div 52 (weeks in a year) = 74 work years

SFW District – Ft. Worth (number of documents based on estimates from the USACE)

- $13,130,880 (8 \frac{1}{2} \times 11 \text{ and } 8 \frac{1}{2} \times 14 \text{ pages}) + 547,120 (\text{E \& C sized pages}) = 13,678,000$
(total # of documents at district location)
- $13,678,000 \times 30$ (seconds for each document to manually enter metadata) = 410,340,000
(total seconds for all documents to be manually indexed)
- $410,340,000$ (total seconds) $\div 144,000$ (seconds in a work week) = 2,849.58 weeks
- $2,849.58$ (total weeks) $\div 52$ (weeks in a year) = 54.80 work years

NWS District – Seattle (number of documents based on estimates from the USACE)

- $10,402,560 (8 \frac{1}{2} \times 11 \text{ and } 8 \frac{1}{2} \times 14 \text{ pages}) + 433,440 (\text{E \& C sized pages}) = 10,836,000$
(total # of documents at district location)
- $10,836,000 \times 30$ (seconds for each document to manually enter metadata) = 325,080,000
(total seconds for all documents to be manually indexed)
- $325,080,000$ (total seconds) $\div 144,000$ (seconds in a work week) = 2,257.50 weeks
- $2,257.50$ (total weeks) $\div 52$ (week in a year) = 43.41 work years

Clearly, a solution involving automated or at least partly automated metadata collection can save the Corps tremendous time and expense. For example, technology available in today's best-of-breed document capture software is capable of automatically recognizing and entering form data as metadata.

4.3.7 Information Reuse

Many software applications today involve tasks that require the sharing of geographic, database, financial, and document information. Interoperable solutions provide end users with a means to retrieve and share data. Creating a solution for data and knowledge exchange among users of information systems presents several questions, among them:

- What data is available for reuse?
- Who will need to reuse available data?
- Who will be permitted to reuse the data?

These questions and many others arise when planning to implement information systems that enable users to share information across an enterprise-wide system.

In the Corps today, information accessibility is limited to the bounds of the department walls. The reusability of archived paper information is limited: information contained in boxes located in distant warehouses is very difficult to share and reuse, and poses direct limits even on what data can be collected. Electronic storage formats for files can also present limits to viewing, and thus reusing, information. For example, CAD application vendors require use of a licensed viewer or application merely to view CAD information. This solution generates expenses when

workers need to view an aging CAD drawing, in that electronic format conversion techniques must be created so that standard software tools can be used to view CAD drawings.

The proper reuse of information will definitely save time and money in the Corps. The challenge in the Corps is to plan to implementation a system that will not only store scanned electronic data, but that will also be capable of sharing that stored data across Corps districts, divisions, and departments. An EDMS or EDMS-like solution has the potential to create new opportunities for Corps workers to reuse and access the valuable library of information generated by the Corps over its history.

4.3.8 Funding

Generally, it appears the Corps considers document conversion an overhead function, and funding is therefore generally not available for standalone document conversion tasks. Often, however, funding is easier to obtain if improving a workflow for the delivery of a project (where the improvement means a reduction in cost or a tightening of schedule) provides document conversion as a byproduct. Therefore, obtaining funding for document conversion requires careful consideration when planning projects.

5 USACE Recommendations

5.1 Conversion Strategy

The Corps conversion strategy consists of two components. The first component focuses on the output product of document conversion, the converted document. The second, the management component, focuses on the management of automated document conversion system acquisitions and the high-level requirements that these systems and services must satisfy.

The conversion strategy must be flexible. Conversion output products must be accessible to multiple applications, unless the functional need dictates otherwise. This flexibility permits reuse of converted documents and is achieved by converting documents to approved technical standard formats.

Document conversion may be accomplished by

- Acquiring specialized but commonly available software and hardware,
- Acquiring document conversion services from another DoD organization, and
- Acquiring document conversion services from the marketplace.

The output product of any conversion must conform to the same flexible technical standard format(s) regardless of the acquisition vehicle used.

Document conversion may be viewed as a two-stage process. Each stage produces a conversion output product that conforms to approved technical standard formats. The product of the first stage results in an interoperable baseline(s) and, therefore, is usable across a very wide variety of application software environments. The first stage output is input to the second stage. The product of the second stage conforms to more specialized technical standard formats. These standard formats allow for additional information (“intelligence”) about the original document that greatly increases its value as a revisable document. However, this often greatly increases the conversion cost.

5.1.1 Strategy Goals

From a technology and acquisition point of view, a conversion strategy should seek to:

- Maximize the flexibility and utility of converted documents to Corps document users through the requirement to use accepted and widespread technical and formatting standards
- Ensure that the output of a conversion effort supports a Corps mission or business process in a cost-effective manner
- Avoid unnecessary duplication in the acquisition of document conversion hardware and/or software.

5.1.2 Document Conversion Activity

A document is a type of record. Historically, the original document medium has typically been paper, microform, or another non-digital form of storage. Document conversion is the process or activity of moving the information content of a document to a digital format (the converted document) while ensuring preservation of the information. ADCS strategies enhance this process by using imaging technology to perform the conversion with minimal human intervention.

Digital documents have distinct advantages over analog documents (*e.g.*, paper and microform documents). Digital information may be processed by machine, may be shared through communications links, and in most cases, may be less expensive to locate, retrieve, manipulate, and disseminate.

All ADCS procedures should apply consistent concepts of document manipulation, including preparation, indexing, scanning, quality control, storage, and retrieval of the converted document throughout the conversion process. The process for the “CONVERT DOCUMENT” activity reads as follows:

- The Convert Document activity transforms the Original Document into a Converted Document. Non-Converted Document and Source Document are other outputs. This activity uses Cost Analysis Tools, ADCS Tools, Human Resources, and Facilities (the mechanisms), under the constraints of ADCS Policy and Budget.

Two-Stage Document Conversion Process

Digital documents conforming to standard formats are more easily shared among functional applications. Sharable digital documents are said to be in an interoperable format. Digital documents must not be locked into a proprietary digital format.

In recognition of the cost and long lead times associated with large conversion projects, changing technology, and the potential multiple uses of digital documents, document conversion activities can be finalized at one of two stages of completeness:

Stage 1: Interoperable Baseline

The first stage establishes an interoperable baseline of digitized information. Original documents are converted to a non-proprietary, interoperable format, allowing digital documents to be shared by multiple software applications on different computing platforms.

This first stage conversion process is often labor intensive, because existing paper files must be inventoried, packaged, and scanned. The use of standard, reusable methodologies, procedures, and equipment will minimize these costs. There may need to be more than one interoperable baseline, based on identified document types. The baseline format may be a raster standard for images, or American Standards Code for Information Interchange (ASCII) for text.

Stage 2: Intelligent Document Forms

The second stage establishes intelligent document forms that are not satisfied by the interoperable baseline but are required to satisfy mission- or business-justified requirements.

A goal of the second stage is to establish standards for intelligent formats; formats that permit easy document revision. However, a transition period will occur during which proprietary formats may continue in certain applications. The enhanced format may be a text format such as Standard Generalized Markup Language (SGML), or a vector-type standard for engineering drawings. In addition, the emerging standard Extensible Markup Language (XML), positioned as a possible successor to SGML, should not be overlooked.

Benefits of the Process

This flexible two-stage approach extends the potential for reuse of a converted document to satisfy different user requirements for the same document and to develop advanced, intelligent, standards-based converted document formats required to satisfy more demanding mission- and business-justified needs.

First-stage conversion ensures that the information contained in a converted document will be accessible throughout its life, as specified by business needs. The life of the converted document is the same as the original record.

For some applications, it is cost-effective to integrate both stages (first stage and second stage) of conversion into the same conversion procedure. However, such integration may be counter-productive in the long term. For example, a National Archives and Records Administration (NARA) study examined several conversion programs. The study states that documents converted without a separation between first stage and second stage may be locked into a proprietary format. This situation may cause the converted document to be unusable by other functional applications. Unlocking the information in these documents may require a second conversion, possibly from the original document.

Standardization is a contributor to success of the interoperable baseline approach. For example, the Defense Printing Service (DPS), in conjunction with the Congressionally sponsored ADCS Test, demonstrated the capability to convert baseline documents to application-oriented documents.

Documents that have undergone first-stage conversion become an interoperable baseline from which information can be extracted to meet varying functional requirements. These documents are in a non-proprietary format and are identified through uniform indexing rules. For example, converted paper and microform-based document formats may be prescribed in MIL-R-28002B. On the other hand, formats such as aerial photography may require a different standard baseline format. Thus, multiple baselines may exist to satisfy unique functional requirements. It is imperative that any conversion baseline be non-proprietary in format with a uniform index.

Integrating both conversion stages is acceptable if the data format created from the second stage is non-proprietary and interoperable; or if the system for which the data is converted easily converts the data to a non-proprietary and interoperable format.

5.1.3 Business Case Review Activity

This section provides additional information on how to conduct a business case review based on the conversion strategy. It begins with an analysis of the requirements and an assessment of the mission or business environment to provide the information needed to support the business case.

It then addresses issues involved in developing a cost justification or economic analysis. Methods for assessing technical capability of ADCS system architectures are also introduced. Finally, this section introduces the key elements of a decision table to assist program managers in determining if ADCS for their specific application will meet their operational need and produce sufficient cost savings or cost avoidance to justify the conversion.

Document conversion is a managed activity. The acquisition of equipment, procurement of services, development of technologies, and development of information systems support document conversion. Consequently, such acquisitions must:

- Be justified by a business case
- Conform to Federal Information Processing (FIP) resource and DoD automated information system management policy

A document conversion program should be considered:

- Only when a military mission or business reason exists for conversion. For collections that include record and non-record material, convert non-record material
- (Non-relevant documents) only if the cost of separating it from record material exceeds the cost of conversion.
- When document conversion is more cost-effective than document re-creation. Convert organization documents of poor quality only if the cost to convert the document contents through Optical Character Recognition (OCR) is lower than direct re-keying. The cost of scanning and correction of OCR errors may be significantly higher than direct re-keying.
- Only when the converted document will be operationally usable in the intended application(s). Conversion programs should focus on documents in which the conversion or end-user system provides adequate image enhancement and editing capabilities to restore any information lost during scanning.
- When the converted document may be accessed by multiple applications. Conformance to standard formats will ensure that multiple applications can access valuable business documents directly or indirectly (*i.e.*, through translators or “bridges”).
- When the long-term cost of maintaining or using the document in its original medium exceeds the long-term cost of maintaining the document in a digital medium.
- When the costs of migrating a document over its life include consideration of the long-term stability of the proposed storage medium and the long-term integrity of the storage medium.

5.1.3.1 Funding

Generally, it appears the Corps considers document conversion an overhead function, and funding is therefore generally not available for standalone document conversion tasks. Often, however, funding is easier to obtain if improving a workflow for the delivery of a project (where the improvement means a reduction in cost or a tightening of schedule) provides document

conversion as a byproduct. Therefore, obtaining funding for document conversion requires careful consideration when planning projects.

5.1.3.2 Requirements Determination

An automated document conversion project must be viewed in the context of its mission or business environment. This is accomplished during the development of the business case in support of the requirement. The requirements (mission need) definition must establish a clear justification for conversion. Cost and performance measures will be developed to evaluate conversion projects.

As a necessary pre-condition to a conversion project, the justification must be approved and meet the following tests:

- *Need for information contained in the original documents*

No conversion should be undertaken unless the information contained in the original document(s) is relevant to the supported mission or business function. For example, conversion of technical documents should be limited to original documents that support currently maintained weapons systems.

- *Demonstrated/quantified improvement in the military mission or business process resulting from converting original documents*

This assessment must be made in the context of other considerations, such as the need to protect the original document from permanent loss or the need through document conversion to accelerate an overall business process. For example, the overall cost resulting from a delay in corrective maintenance caused by inadequate access to technical documents may exceed the cost savings realized by not converting technical documents with a low probability of access.

5.1.3.3 Information Access and/or Dissemination

The choice of indexing schemes and data elements determines accessibility. The use of a uniform indexing scheme, selected to optimize retrieval requirements of converted documents throughout DoD, is essential to the design of consistent interfaces between the conversion system and any end-user systems. Also essential is the identification of local repositories of converted documents. This is the responsibility of the functional official. The adoption of standard data formats further ensures consistency among converted documents.

5.1.3.4 Archived Converted Document

The converted document should be stored in an “archival” form to ensure that a converted document, throughout its life, remains available to appropriate applications. Because technology is changing quickly, there is no guarantee that future digital systems may be able to read recordings made on older systems, even if these recordings are still in good condition. In addition, there is no guarantee that *de facto* standards will persist during the life cycle of either a given system or the converted document produced and stored in that system.

However, conversion projects will need to make provisions for the delivery of digital documents on an appropriate medium that is in the mainstream of technology. The requirement upon future archives must be limited to only those data and media that are certain to be supported in the out-years. Conversion projects must also provide for the safekeeping of conversion software and hardware in sufficient quantity to support document access or future re-conversion to maintain accessibility, until scheduled document disposition.

The requirements for archiving also need to make provisions for back up of the converted documents and disposition of the original documents, in accordance with DoD Records Management policies and procedures. For example, NARA will accept, for permanent records in electronic form, only “one-half inch, seven- or nine-track reel-to-reel magnetic tape and 3480-class tape cartridges” and “CD-ROMs... that are in conformance to ISO 9660 standard.”

5.1.4 Other Issues

The contents of the converted document must remain available regardless of changes in storage technologies, systems, and applications. NARA recommends reconverting or copying converted records to remain compatible with new storage technologies and to satisfy approved disposition requirements for temporary and permanent records.

Consequently, the program manager must articulate:

- How the converted document will be managed until it has reached the end of the scheduled retention period (*e.g.*, conformance to NARA policy)
- How the intended application may impact future re-conversion requirements
- What migration path, such as the adoption of integrated processes and interactive documentation, will be used to accommodate changes in information systems

5.1.4.1 Cost Justification

Requirements for cost justification will include operational, financial, and long-term concerns:

- *Operational.* The operational justification must state that the converted document will support the mission or business requirement and that the information contained in the document cannot be obtained from another source in a more cost-effective manner.
- *Financial.* The financial justification must state the financial impact of the conversion process. Additionally, the justification must indicate whether to contract for conversion services, to buy conversion hardware and software, or to use a centralized DoD conversion service.
- *Long-Term.* The long-term costs associated with managing media for storing electronic documents, hardware to access the media, software to read the media, and migration to newer media due to deterioration of old media must be well understood. Planning should include these costs for the life cycle of the information being managed on electronic media.
- *Technical Capability.* The technical capability of an ADCS system will be determined by an evaluation of the system’s architecture and its conformance to published standards.

5.1.4.2 Business Case Decision Table

The decision table below is provided as a guide to assist program managers in determining if the proposed ADCS acquisition will:

- Meet operational requirements
- Produce sufficient cost savings or avoidance to justify the conversion

Table 5-1 – Business Case Decision Table Requirements Determination

General
Is there a legitimate mission or business need?
Will Records Management requirements be satisfied?
Are the records scheduled per NARA requirements?
Are NARA archival requirements satisfied?
Cost Justification
Can the information contained in the documents be obtained from another source in a cost-effective manner?
Will automated document conversion reduce costs?
Has a comparison been made of purchase vs. contracting the automated document conversion service?
Should centralized DoD conversion services be considered?
Document Candidate Selection
Are the documents active and do they have sufficient volume?
Are the documents available to multiple users?
Do the documents contain valuable and relevant information?
Do the documents have a relatively long active life remaining?
Are the input/information processing routines stable?
Can the original documents be destroyed after conversion?
Technical Capability
Architecture: Does the selected architecture support the minimum functionality, ensuring the interoperability of the converted documents?
Standards: Are relevant standards identified to ensure interoperability of the converted documents?

Once the decision is made to proceed with automated document conversion, the following should be used as conversion strategy guidance:

First Stage Conversion

- Use the DoD standards from Table 5-1 or provide a business case for use of proprietary standards and a migration strategy to open systems.
- Evaluate standards to ensure they meet the interoperability requirements and use non-proprietary formats.
- If first-stage document conversion meets the end user's functional/mission requirements, proceed with LCM documentation. If the conversion does not meet requirements, then move to second stage conversion.

Second Stage Conversion

- Tailor document conversion to meet end-user requirements.

- Evaluate standards to meet interoperability requirements and use non-proprietary formats.
- If interoperability cannot be achieved with second stage conversion, evaluate the cost of maintaining an interoperable baseline copy after first stage conversion.

Functional Officials

The senior functional officials establish policy in their respective functional areas of responsibility. The document conversion role of a functional official is to:

- Implement DoD and/or Corps policy to maintain, protect, and preserve organization records
- Determine the business need for document conversion
- Support efforts to improve the accessibility to functional documents
- Validate the business need and justification for document conversion

Development of automated document conversion functional process and data models is the responsibility of functional officials in accordance with approved guidelines.

Commanders, Managers, Records Officers, and Document Custodians

These officials must:

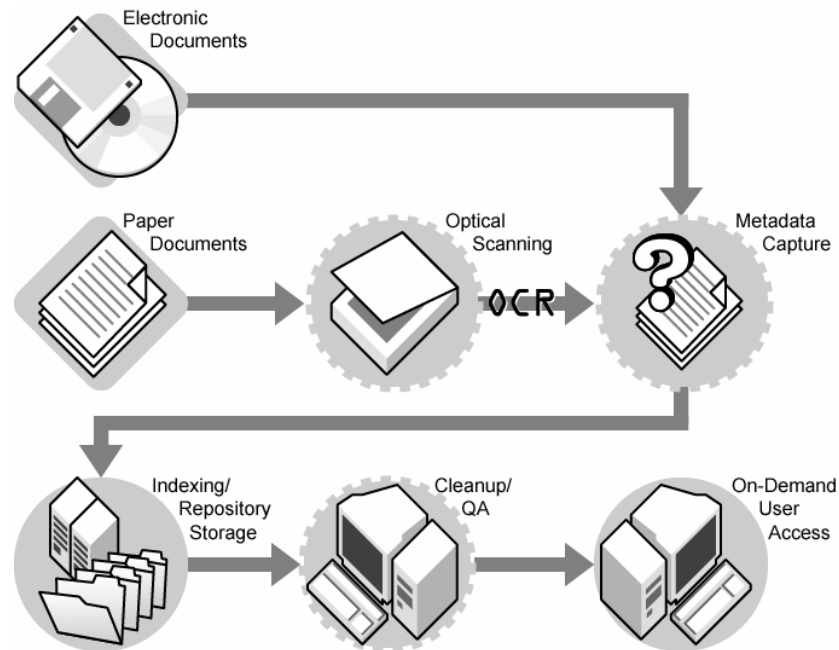
- Coordinate with the functional manager, records officer, systems support officer, and legal officer to determine eligibility of an original document for conversion
- Determine if the original or the converted document will be the organization record, in accordance with the organization's published records disposition schedules approved by the Archivist of the United States.
- Consider the long-term financial costs of managing documents in either their original medium or in their converted medium
- Consider the stability and life-cycle of the proposed medium of storage for converted documents
- Consider the life-cycle of the hardware and software used to store, retrieve, and manage a converted document throughout its life
- Select documents for conversion based on the information value, need to share the information in an electronic environment, and number of times the information will be accessed over its life.

The recommended document conversion strategy for the Corps is divided into five areas of interest:

- Document Capture
- Metadata Capture

- Output to Document Repository
- Access to Converted Documents
- Bulk Loading Mechanism for Future Systems

Figure 5-1 – Recommended Document Capture and Access Strategy



5.2 Document Capture

Document and content capture systems provide a cost-effective, accurate, and operationally simple mechanism to feed content into an EDMS, as well as into other key business applications. Without effective capture of document content and indexable metadata, processes based on accessing legacy data starve and die. How organizations transfer legacy information into present processes and knowledge makes a key contribution to the overall success of the organization.

Document capture is the area where paper documents/drawings are scanned and converted into digital media or where electronically stored file formats are converted to a generic format that can be viewed with standard software utilities. This is the case with drawings stored in AutoCAD® or Microstation™ file formats. Currently, only users who have a license for the correct (not necessarily the most recent) version of AutoCAD® or Microstation™ can view files in these formats. If these files are converted to a more widely used format like TIFF or JPEG, then any user can view these drawings with a standard Web browser or free image viewing utility.

The following areas must be considered when developing a procedure for conversion of legacy data:

- Document Type: Is the document requiring conversion composed primarily of text, graphics, or a combination of both?
- Document Format: Is the document in paper, microfilm, or other digital format? Each format will affect the conversion process and costs associated with conversion.
- Document Use: Once the document is converted to a digital format, will it be used only for viewing and output or will it need updating?
- End Users: Who is the ultimate end user? Will the end users of the digital document be civil authorities, analysts, managers, or those involved in acquisition?
- Distribution Method: Will the converted document be distributed digitally, on-line, or in paper format?
- Update Requirements: How will the document be updated, how often, and when?

5.2.1 Document and Content Capture Defined

Core document and content capture technologies are mature and the understanding as to how to deploy them effectively is well developed. These systems have developed a strong reputation for delivering pragmatic return on investment, as well as for equipping an organization to face doing business electronically.

Document and content capture is defined as the set of technologies and services required to capture documents and information from documents—paper, microfilm, fax, and electronic—in order to process the content of the documents in a form that meets the need of the repository or business application being served. As such, document capture represents a key capability, helping organizations adapt to doing business electronically, and equipping them to harmonize paper-based processes with electronic approaches. It addresses the key issues of:

- Scanning and capture
- Quality Assurance (QA)
- Indexing (Automated and keyed attribute information)
- Output to repository or business applications

The input is a document, whether paper or electronic; the output is an electronic document, metadata, and/or data contained within that document.

5.2.1.1 Capturing Inputs

Incoming content can be considered in four main ways: paper, microfilm, fax, or electronic. If electronic, the emphasis is first on recognizing the digital format of the document, and understanding how to handle it and capture its metadata without re-keying or recognition. If

paper, film, or fax, the document is first scanned to create a digital image, from which one may need to extract part or all of the content using recognition technologies.

5.2.1.2 Processing and Validation

The capture and validation processes to be followed apply equally to both digital and paper content. Depending on application needs, there will be requirements to classify documents into meaningful groups, index them, process any forms, and conduct validation and quality assurance (QA) processes.

For example, incoming post is scanned, Quality Assured, and indexed. An email with a similar content can be printed out and subject to identical processes, or its metadata and content captured digitally and then subject to those processes. The capture process needs to take into account any of the following needs the organization might have:

- **Batch Management of Documents:** Treating a certain set of documents as a group may have procedural benefits.
- **Scanner Management and Control:** Uniform requirements must be imposed for determining scanning resolution, types of images required, image compression, image processing and enhancement, and others.
- **Types of Scanners:** Today, scanners are available to fulfill a broad range of needs, from hand-held text capture devices to machines capable of handling 4,000 items per minute; from digital copiers with scanning capability to high-power microfilm scanners; from credit card scanners to devices for capturing A to E-sized engineering drawings and maps — all in monochrome or full color, single or double-sided. As such, it is possible to match the scanner — or mix of scanners — closely to the inputs.
- **Determining Output Formats:** Digital output formats depend on: (a) how the document is to be stored, used, and viewed; and (b) whether it is to be retained in its original, usually revisable format (*e.g.*, Word or Excel), or rendered into a non-revisable format (*e.g.*, Adobe PDF). Such decisions depend upon the application.
- **Quality Assurance:** Scanner monitoring, batch management, and image quality checking functions are important to ensuring the usability of the final product.

5.2.1.3 Output from Document Capture

Output can be transferred into a business process or into a repository—or to both (very common).

- **Output to a business process application:** If output is to a business process, then decisions need to be made as to just how that is to be achieved — as part of a workflow system, through proprietary links, or via some Enterprise Applications Integration (EAI) middleware link. These need to be capable of passing the documents and/or data to one or more linked applications as required.

- Output to a repository application: Where the document is to be captured in an EDMS, managed as a business record, or eventually archived, then best practice is to enter it into a repository, in addition to any processes it may be supporting.

5.2.2 Vectorization

Technology has improved significantly to enable more intelligent documents with more sophisticated interactive linking of supporting data. For example, Commercial-Off-the-Shelf (COTS) products for organic raster to vector conversion and intelligent text conversion (see below) are readily available in the marketplace. Many of these products are easily integrated into the existing environments and are intuitive enough or are extremely user-friendly that formal training is not always required. However, not all products are homogeneous in support or capability. Therefore users must ensure that the conversion software used is compatible with their existing environment.

While use of vectorization tools can be advantageous to many processes, a careful examination of appropriate levels of data forms is required. Determining the actual use of the information will provide data managers with knowledge of appropriate levels of conversion. Vectorization provides an intelligent two-dimensional attribution of data; however, a three-dimensional product model is often required to allow for automated re-tooling of vector data.

Accommodating the Corps' business needs dictates that a "menu" of conversion alternatives be made available to support business requirements as well as the financial constraints of reduced funding levels. Six drawing conversion levels, from simple to complex, have been defined. It should be noted that each of these format levels has a particular place in a conversion strategy. The format level of conversion is primarily based on the ultimate use of the data and the relative costs associated with the level of conversion. These conversion format levels include:

- Level 1 - Raster Image: A scan of the original drawing which converts the drawing into a raster image, also called a bitmap. It is the lowest cost alternative and produces a very large file size, which is generally non-editable.
- Level 2 - Raster Image plus Cleanup: Essentially Level 1 enhanced by cleanup and de-skewing. It removes unwanted entities from the originals and squares the drawing. The file is reduced in size and the image clarity is improved.
- Level 3 - Automatic Vectorization: Converts the drawing to vector representations of the raster scan, which are in files that can be loaded into a CAD system for editing.
- Level 4 - Auto Vectorization Plus Text: Essentially Level 3 enhanced with ASCII text replacing the automatically vectorized text. The text and dimensions are now recognized as text rather than vector data and are editable. Legibility is significantly improved.
- Level 5 - Enhanced Vectorization: Level 5 adds intelligence to the drawing by cleaning up the vector quality. Circles, arcs, and other entities are true and precise geometries. Lines are continuous and layered. Objects are clear and orthogonally correct.
- Level 6 - CAD Perfect: Level 6 results in a CAD-perfect file. Video tracing or direct CAD redraw are the two processes that produce this level of conversion. All entities are

dimensionally and orthogonally correct with fully editable vectors and text. Layers, blocks, symbols, line types, and current NSI standards are incorporated.

5.2.3 Text Recognition

Recognition technologies are used to extract textual data from documents, for indexing, or for input to a business process. Some important issues to consider when using recognition technologies are recognition rates (how much can the technology recognize) and confidence levels (how accurate is the data and metadata captured). The Corps should consider text recognition technologies for two reasons:

1. Successful text recognition can convert printed characters into text-aware documents (word processing documents in particular).
2. Advanced text recognition can partly automate metadata collection, significantly reducing the time it takes to properly index the scanned documents.

Recognition technologies include two basic varieties: text and document recognition.

5.2.3.1 Text Recognition Technologies

The bulk of text recognition technology centers around character recognition: turning printed or written material into a text-based electronic file.

Optical Character Recognition (OCR)

OCR is generally used for recognizing machine-printed characters of good quality and across a set range of fonts, point sizes, and type weights. This gives the best results, in many cases, but the process relies on the OCR system's existing knowledge of the fonts used, so OCR is somewhat inflexible.

Intelligent Character Recognition (ICR)

ICR is used for recognizing poorer quality machine print, or where the font is not predefined, and constrained handprint (the reading of handwritten block capitals and numbers). ICR also includes a capability to "learn" fonts and handwriting styles in one of two ways.

- Feeding the ICR system a representative sample of the types of documents to be processed and causing it to "learn" from the way an operator corrects recognition errors.
- Working with an ICR system that progressively "learns" from the documents that pass through it and from the manual correction they require.

Some recognition systems include cursive (joined-up) handwriting recognition, currently used mainly for reading addresses on envelopes and the written amount on checks. As the technology improves, it will become possible to recognize handwriting in other contexts.

Magnetic Ink Character Recognition (MICR)

MICR is used mainly on bank-issued checks, to read printed account information, check numbers, and sort code details.

Bar Code Recognition (BCR)

BCR is used to translate “thick-thin” vertical bar patterns into a numerical sequence, usually for inventory control or point-of-sale purposes.

Optical Mark Recognition (OMR)

OMR is a technique for recognizing pre-defined shapes in pre-determined positions – for example a check or a cross in a box. It is used in applications such as standardized testing, where a 100% accuracy rate is (hopefully) achieved through simple box marking. Software verification can be used to detect errors (such as checking too many boxes), or even to distinguish between a checked box and one that has been scrubbed out. OMR can also be used for signature detection – not to recognize the signature, but to check that “something” has been entered in a signature box.

Figure 5-2 – Character Recognition Technologies



5.2.3.2 Document Recognition Technologies

Optical Document Recognition (ODR)

ODR, also known as template-based forms processing, is used to process paper forms by recognizing and extracting data from them in order to feed the data into a business application. ODR is particularly appropriate for survey processing, through government forms to mail order purchasing and many others. Any application involving structured forms completed by machine or by constrained handprint is a good candidate for an ODR forms processing approach.

ODR works by understanding defined forms templates. These tell the system how to identify a particular form, which fields need to be recognized, what types of field they are, and what rules can be applied to maximize recognition results.

There are a wide range of techniques employed within this technology, including:

- Customizing image clean-up and validation rules.
- Making more sophisticated use of recognition engines.

- Using multiple engines to select the “best fit” result.
- Passing the image through the recognition engine multiple times at different image settings to optimize the results.

Once the core data has been recognized, any documents that have failed are routed to reject repair operators, who check the forms and re-key any incorrect data. Re-keying is usually done with reference to the actual document image, which is presented to the operator for this purpose. Key issues critical to the success of such systems are:

- Ease of template definition and/or forms design
- Quality of recognition
- Avoiding “false positives” tends to be more important than doing additional reject repair but that depends on the application.
- The management capabilities built into the system
- Includes routing captured forms through the process, monitoring the results, reporting on problems, and managing the scalability of the systems to handle multiple scanning points, recognition stations, reject repair routing, and finally, output of the appropriate data stream to the relevant application.

Intelligent Document Recognition (IDR)

IDR, an emerging technology, appeared first in Europe and seeks to automatically identify a document’s type and its related attributes from the layout and structure of the document. The key benefit of this approach is a significant reduction in the labor hours needed to index documents.

IDR is finding applications in invoice processing and is being deployed for mailroom processing, insurance, and medical applications, among others. The technology can be made to work on old image repositories, which can be “mined” either to export the data into knowledge and Customer Relationship Management (CRM)-type applications or, more prosaically, to re-index the documents so they can be re-used for new purposes brought about as a result of change in the business, through merger, acquisitions, business refocusing, etc.

5.3 Metadata Capture

Metadata capture is the collection of information about the document. This step deserves dedicated effort to add critical value to the converted documents. Though metadata capture can involve significant time investment, and takes a trained and knowledgeable individual to perform correctly, collecting all the appropriate data about the document or drawing being converted is essential to the purpose of converting the documents in the first place.

Indexing (Capturing Attribute Information)

Metadata capture is of primary importance when contemplating a document conversion process. The concepts behind metadata have been used by librarians to generate card catalogs for generations (only recently have these been computerized). For example, metadata items relating

to library books might include the book's title, author, subject, and Dewey decimal number. A list of suggested metadata items appears in Appendix B. A document's type is one important item of metadata; a list of suggested document types appears in Appendix C.

Information about a document's content brings the power of a database to electronic document repositories. Just as you could search for an item using the title, author, or subject in the library card catalog, you can search for electronic documents using an unlimited number of fields, or meta tags. Although users cannot see metadata when opening files using an information browser, metadata remains associated with a file no matter where the file moves. Metadata works with any file format and media type, requires minimal user interaction, and employs universal bibliographic standards.

5.4 Document Repository

After a document is converted to a reusable piece of digital media, the information should be then stored in a digital storage location like a file server. At the Corps sites surveyed, subcontractors deliver scanned paper documents and drawings on CD-ROM. While this is a marginally usable temporary storage system for a single user, such a system has several drawbacks.

- Searching a stack of CD-ROM disks for any particular file is difficult and time consuming.
- Accessing a CD-ROM is impossible for more than one simultaneous user without significant investments in special equipment, user training, and networking capacity.
- Indexing the CD-ROM tends to reveal only perfunctory metadata, if any.
- Damaging the CD-ROM renders its contents unusable and lost.

(See also Section 4.3.4.) Converted documents and drawings must instead be stored more permanently in an accessible way. For optimal accessibility and safety, a document repository should be established on a networked file server. Converted documents can then be made available to all networked users, and the data is then regularly backed up to tape or other backup media according to the department's backup schedule.

To ensure digital data can be properly accessed and used after conversion takes place, other significant factors must be taken into consideration. The following are essential factors in determining the end use of digital data:

- **Integration:** What are the infrastructure requirements? How and where will the converted documents be stored? What systems required interfacing and what method of access will be used?
- **Storage:** What level of indexing will be required for document management? What are the storage requirements with regard to sizing, location, and backup?
- **Maintenance:** How will the documents be maintained? Who is responsible for update? Will multiple output formats be required?

- Output: What formats must the converted document be in? What media must be used and how will it be accessed?
- Tracking Systems: Will the conversion system be interfaced to the customer? What standards must be met? What criteria will be used for ordering and estimating the conversion process? How will the status of the conversion process be monitored and reported?

5.5 Document Access

5.5.1 Document Management

There is a need in most design organizations to reduce the high labor cost of locating and accessing data. It's been said that a large design organization spends about 30 percent of an average technical employee's time just trying to find documents and other information. If this number is valid, it represents a considerable cost. For medium to large design organizations, the implementation of a design- and CADD-focused electronic document management system (EDMS) can produce considerable net cost savings. To realize this benefit, much thought and planning must go into predicting how future users will go about the process of searching for and retrieving files and documents.

The definition and capture of metadata (data about the primary data being stored) becomes of critical importance in large organizations, particularly for CADD-using design organizations. The metadata will provide the ability for users to search for and locate information stored within the system. For large organizations like the Corps, the standardization of metadata is critical for long-term ability to locate the original documents or files.

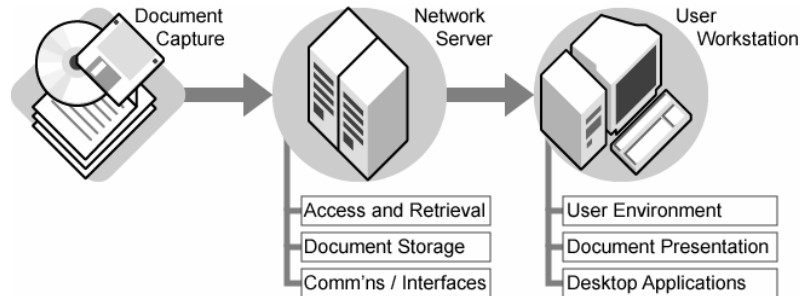
As an example, one recent project had a need for an electronic method of storing drawings that were rapidly becoming unusable due to deterioration. The drawings had been stored in sets of flat files in numbered drawers. Without putting much thought into the process, the conversion effort was contracted out to an outside vendor. The drawings were scanned and the resulting files were named "00001.tif," "00002.tif," etc. Indexing the files for searching was required, and the metadata stored for each file was very simple – the filename, the stack (set) of flat files, and the drawer number. This system was acceptable for the existing staff that had worked on the project for years – those staff members could locate the necessary file in only a few tries because they had been finding the drawings for years in the flat file drawers. But this "remembered" metadata will be lost when the staff members with those memories retire. Protracted, almost blind, searching will be required in the future.

For long-term projects, it is also important to deal with the records management and migration issues for CADD, GIS, and even word processing data. In some organizations, projects have spanned 20 to 40 years and three to four generations of computers and software. In many instances, the project data has been re-created in new formats (rather than migrated) several times, at considerable expense. This cost can be minimized in the future through proper life-cycle records management and planning.

In the area of CADD-based drawing production, management of CADD files during production is critical to productivity. Much of the automation that can produce significant reductions in

labor and schedule require that a program be able to resolve the location and status of files on its own.

Figure 5-3 – Recommended Converted Document Access Scheme



5.5.2 Planning for Digital Permanence

Reports started appearing in the press in the late 1990s:

- Magnetic tape containing 1970s-era satellite photo survey data of the Brazilian Amazon cannot be accessed to establish deforestation trends.
- Twenty percent of the data collected during the 1976 Viking Mars mission can no longer be read.
- In the State of Oregon, the primary database of people with disabilities vanished.
- Some prisoner of war (POW) and missing in action (MIA) records and casualty counts from the Vietnam War can no longer be read.
- At Pennsylvania State University, all but 14 of some 3,000 computer files containing student records and school history are no longer accessible.
- The paper recording the United Kingdom's 1911 census now crumbles when touched.

Much more has probably been lost. The problem is that gaps in digital records are undetectable until someone tries – and fails – to retrieve information, by which time it is too late. There is little assurance that today's information will retain any more value than a VisiCalc file created in the 1980s and stored on an eight-inch floppy disk: inaccessible information is lost information. Retaining and maintaining accessibility to digital information requires constant human attention. The following issues are vital to a well-thought-out preservation strategy:

- *Longevity* – The ability to access and read digital documents in the future with virtually no degradation in information content, including formatting.
- *Inter-operability* – The ability of the digital preservation technology to work with existing and future IT systems to provide access to the preserved documents.
- *Total Cost of Ownership* – A lower total lifecycle cost viewed over the long term, including capital expenditures, storage media, operational expenses, maintenance contracts, and migration/conversion costs.

- *Technology Obsolescence Protection* – The ability to provide access to facts contained within digital documents through successive generations of new software and hardware.
- *Backup and Recovery Support* – The ability to support business functions, if necessary, using preserved documents.

Emerging digital preservation technology is offering a way to deliver permanence while addressing all of these issues. This is welcome news for document-intensive organizations with longer-term retention requirements such as insurance companies, financial institutions, health care centers, hospitals, government agencies, and public utilities.

5.6 Bulk Loading

“Bulk loading” is the process of taking up a known set of data and transferring that data into another system, database, or file. The bulk loading process can also include operating on the data before transfer; this is sometimes referred to as “data transformation.” Data transformation is required when the system into which the data will be loaded requires that a data set meet certain requirements before reuse. Bulk loading includes two types of solutions. Basic replication moves information between databases that maintain the same basic information scheme for all source and target databases. Replication with transformation, on the other hand, makes it possible to move information between many different types of databases, including competing brands and models (relational, object-oriented, and multi-dimensional), by using translational routines to transform the data so it is represented correctly to the target database or system.

Bulk loading provides a way of providing a large set of data at the beginning of a software implementation. Bulk loading mechanisms are useful when scanning software applications do not provide mechanisms to transfer metadata into a new information management system or database (most do not). Most bulk loading mechanisms are largely custom-built to meet a specific set of requirements for maintaining links between metadata and documents, for example in an EDMS.

Most state of the art COTS scanning solutions do offer metadata capture. After scanning a document, the user is forced to enter the metadata for the document. Metadata collection becomes a part of the process and not left for later, and consequently skipped entirely. Improvements in zone OCR and pattern recognition techniques give many scanning solutions the capability to not only “read” a document but also automatically populate some metadata attributes. These types of solutions do require that documents follow a certain format or at least have the same data in the same place from one document to another.

When moving this data to new information management systems, however, there is a need to move all the collected metadata into a database not contemplated by the original process. Many database systems on the market have internal import/export facilities: when information is needed from a database, a user might export the database into a common format such as a “comma-delimited” file. The comma-delimited file is then used to import data into a target database. This works well when both the source and target databases are the same, but if they are different data cannot be transferred in this way. In this case, the database must be transformed using a custom “bulk loading” application to move the data from its original database, transform

the data into a format expected by the new database application, and then deliver the transformed data into the target database.

5.7 DCRDs

In years past, Intergraph Solutions Group personnel have been involved in developing Document Conversion Requirement Documents (DCRD) for various Department of Defense agencies. A DCRD is a document that specifies the conversion process for a particular type of document or drawing into a reusable defined media. A DCRD generally describes the probable future need of the documents, the preparation of documents before scanning, the file formats required for output, the details of the scanning procedure, the details of the format quality, the standards for filename convention, and the degree of adherence to CAD standards if applicable. The DCRD is generally used to describe the process of scanning drawings into a raster format and then taking that format into a vector format to be used in a CAD environment.

Conversion requirements for each system are unique and the opportunities for cost-effective application vary among contractor and Government activities. It is essential to have coherent strategy for engineering related document management and to be able to communicate that strategy effectively. Proper conversion to a vector format is a valuable asset that, if managed properly, can save time, increase efficiency, improve system quality and performance, and reduce cost. Decisions must be based on the overall program conversion strategy, defense system support strategy, available technology, and analysis of costs and benefits of alternative methodologies.

For the foreseeable future, document conversion infrastructure must accommodate both legacy and standard processes. An effective DCRD should address:

- Document users, types of documents, frequency of use and timeliness of document access or delivery to each user.
- Government hardware and software systems to use or in development to manage and use converted document data.
- CAD/CAM exchange requirements including format, media, applicable standards, and existing telecommunications/network capabilities.
- Concurrent access by multiple functional users.
- Rights access authorizations and restrictions.
- Document management responsibilities.
- Flow of converted document/data among government site.
- Identification of converted documents and metadata requirements.
- Methods to be used for the exchange and final disposition of document/data

Funding was not available to write a specific, detailed DCRD for every document type found within the Corps. Also, because every Corps District operates semi-autonomously, detailed

DCRDs could never apply to all Corps sites. For these reasons, Intergraph Solutions Group has defined a generic methodology or approach for document conversion within the Corps. In addition, we have provided an example DCRD, which appears in Appendix H.

Prior to beginning any drawing conversion effort, it is imperative that the customer and the conversion contractor adopt a detailed set of conversion requirements. A DCRD such as the Appendix H example serves as the foundation and the guide by which all converted data will conform to throughout the project. If possible, this DCRD should be drafted together and agreed to by both the conversion contractor and the customer. The DCRD will become a “Living Document” that will most likely require change over the course of the effort. Encountering variations in source drawings that are provided for conversion will drive these DCRD changes. Any changes made to the DCRD must be documented in the change section of the DCRD and agreed to by all parties.

The DCRD will outline the events and procedures of the entire effort. This includes, but is not limited to the following types of information:

1. Source Data – This will describe the type and quantities of source data provided by the customer for conversion. Examples could include hardcopy engineering drawings (paper, Mylar[®], etc), technical manuals, existing digital files needing conversion to other formats.
2. Required Output File Types – This would include required file types such as DWG, DXF, Cals Type 1, PDF, XML, specific resolution of delivered raster files, metadata files.
3. Delivery Specifications – Identifies delivery media (CD-ROM), delivery quantities, labels, delivery schedule, delivery addresses and contacts.
4. Quality Assurance & Quality Control Requirements – Typically this identifies the type and amount of Quality inspection each converted file shall receive. It also identifies the procedures to be followed in the event that drawing redlines are encountered.
5. Special Instructions Textual Data – This would include instructions for completing title blocks, notes fields, revision blocks, zones, etc.
6. Data Sensitivity or Security Requirements – Identifies any security clearances or permissions that conversion personnel must obtain before being allowed to work with the customer provided data.
7. CAD Drawing Requirements – Identifies layers, colors, fonts, text size, line-work symbology, etc. of converted digital CAD files.
8. Dimension Specifications- Identifies the types of Dimensions (Associative), Dimension text size and fonts, Converting Fractions, Leader Lines, etc.
9. Metadata or Index File Requirements – Ideally all delivered data will include captured metadata. See Appendix B for typical types of metadata information.
10. Points of Contact, for both the customer and conversion vendor, for the effort.
11. A listing of all Reference documents which are relevant to the effort, including existing Government specifications. There may be instances during the conversion effort where the

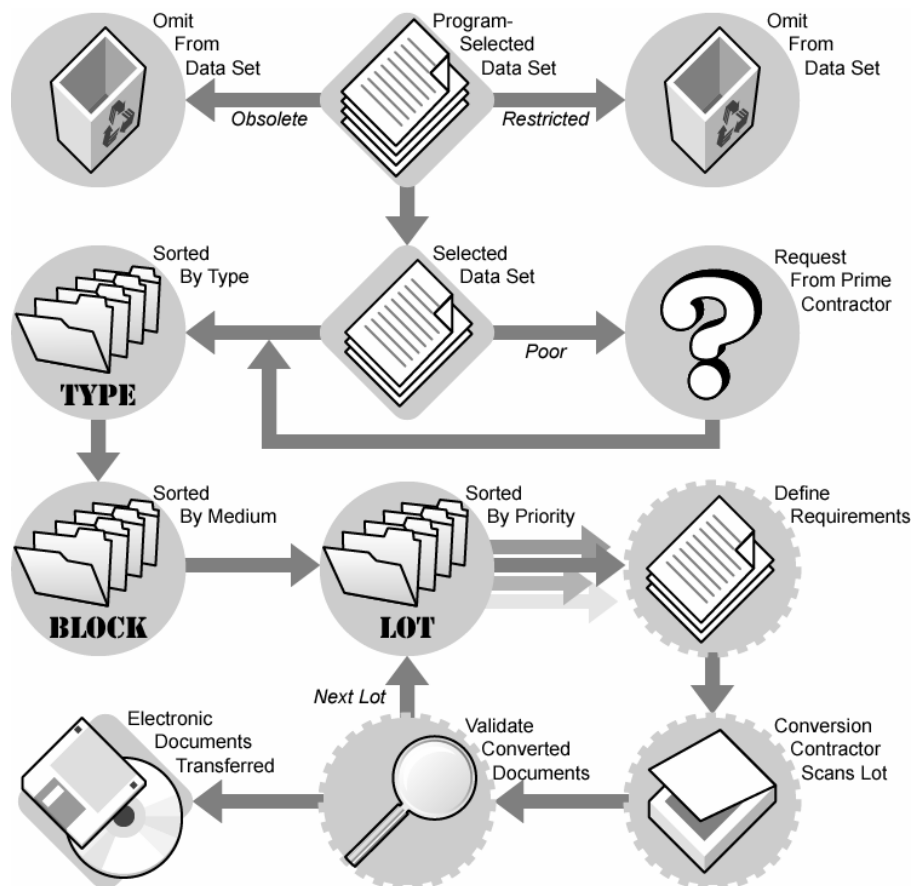
Government specification and the DCRD would be in conflict. A general rule is that the Government specification would take precedence, but all parties should agree this to prior to beginning the effort.

Establishing a mutually agreed upon DCRD prevents many conversion problems. However, there will be instances in which the DCRD will not provide for unique data occurrences. Such instances must be handled on a case-by-case basis between the POCs and all agreements documented. If warranted, these conditions should be added to the DCRD as change pages discussed earlier.

5.8 Conversion Procedure

A documented, well-defined procedure such as that contained in a DCRD is integral to document conversion program planning and management. The Program Manager uses the DCRD to provide strategy for acquisition and use of data throughout a program's life cycle. Comprehensive DCRD development provides the framework for government and contractor process development, documentation, and implementation.

Figure 5-4 – Conversion Decision Matrix



Program Managers must provide effective management throughout the conversion process. A top-level conversion process, as depicted in Figure 5-4, proceeds as follows.

- Identify the program-selected data set, omitting documents that are obsolete.
- Assess rights information pertaining to the data set; any documents the program does not have the right to convert cannot be converted.
- Evaluate the data set for image quality and retention value. In the case of documents that cannot be read, the prime contractor may be contacted if feasible to provide documents of greater quality.
- The number, size, and types of the documents requiring conversion are determined. The documents are first separated according to document type (*i.e.*, proposal, aerial photograph, map/chart, etc.).
- Every document type is further separated into blocks according to format (*i.e.*, paper, Mylar®, aperture cards, microfiche, TIFF, CALS Type 1). For example, within a set of mechanical drawings, there may be paper drawings and there may also be Mylar® drawings. The Program Manager would separate this example set of documents into two blocks: one block for mechanical paper drawings and the other block for mechanical Mylar® drawings. Each document block will have separate conversion requirements and will require a separate DCRD.
- Each block is then divided by priority into lots; document lots with greater priority are scanned and converted first.
- After dividing the document blocks into lots, the conversion requirements for each lot are prescribed according to the DCRD. Usually, all the lots within a single block will carry identical conversion requirements and all the documents within a single lot will carry those same requirements. A sample requirements worksheet appears in Appendix E.
- The output format is determined, followed by selection of the conversion vendor.
- Conversion begins with the first lot. Documents are validated as conversion is completed. Any required changes are made and then validated.
- Once the lot has been converted, the Program Manager is responsible for the converted documents' storage.

This general conversion procedure provides a summary of the decision-making process that leads to document conversion. It may be altered to suit the needs of a particular program or office, but this procedure is generally designed to result in manageable workloads and achievable, measurable milestones.

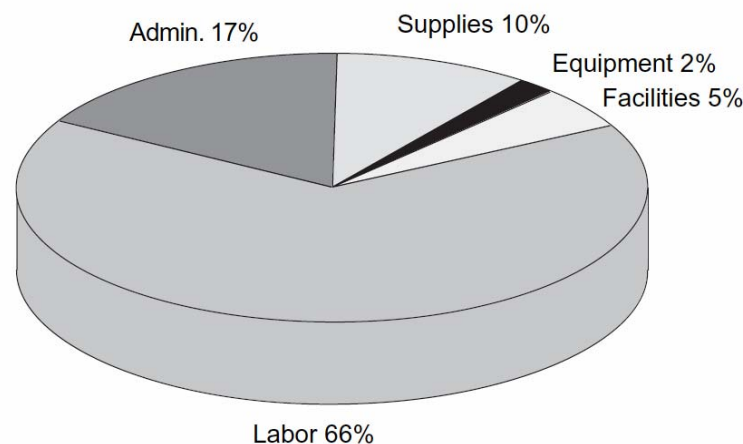
6 Additional Considerations

6.1 Return on Investment/Total Cost of Ownership

6.1.1 Direct Benefits

Scanning drawings to digital form is essential to realize any open archive benefits. The justification for scanning can be easily made when looking at the value of managing the documents, savings in facilities costs, and improved document accessibility throughout the organization. These costs are often not easily measurable and can be staggering.

Figure 6-1 – The Cost of Paper



Paper handling is a major factor in any operation – as much as 30-40% of overall operating cost can be accounted for simply in paper handling activities. A conversion exercise is equally labor-intensive and costly, and compounded by the fact that it is often scheduled over a compressed timescale. Costs to consider include both the overall costs for the conversion exercise and its management overheads, and the cost of “business as usual” during conversion (*e.g.*, disruption to business processes if files are offsite during capture, or of hiring extra staff for an onsite conversion). A carefully planned conversion strategy is vital. Backfiling strategies include full conversion; on-demand conversion; incremental conversion; or a “from this day forward” strategy.

Staggering Costs of Documents*

- 4 trillion documents are stored in US files
- Paper files are doubling every 3.5 years
- Average document is copied 19 times
- Each day, almost one billion photocopies are made

- Average worker has a 34 hour paper backlog
- Half of an office workers time is spent handling paper or data entry
- 50% of all projects are behind schedule

* Source: AIIM, Forrester, Star Securities

Post capture costs must also be considered. The paper may have to be retained for legal reasons (for example deeds and policies), with a cross referencing system to ensure that it can be traced back through the digital record. Even if it is to be destroyed, there may be special controls and safeguards to be applied.

Significant attention should also be paid to weeding out documents before conversion – duplicate documents, superfluous notes, documents that are irrelevant or have been superseded. It is not unknown for organizations to discover that more than 50% of their archive is more suited to shredding than conversion.

Any evaluation of the return on investment for document capture and conversion revolves around the value of the documents that actually should be converted and the processes they support. Some documents have little intrinsic value in themselves, *e.g.*, a memo on a scrap of paper, but their content can be extremely valuable if it records a decision, event, or transaction. Other documents are structured and perform specific tasks, *e.g.*, checks, which contain numeric and free-form text. Others, such as photographs and representations of fine art, engineering drawings, etc., have value embedded in non-character-based content, which is easily assimilated by eye but rarely by machines.

As document capture handles all types of documents, the value can be categorized in three key ways:

- *Quantifiable Cost Savings*: Enterprises are already trimming costs by eliminating expensive real estate used to store paper filing cabinets and replacing their files with data on their servers.
- *Indirect Savings*: Indirect savings are realized whenever a process increases the speed of the business cycle, or “business velocity.” With better access to legacy data, Corps workers will be able to significantly reduce delays, completing work faster.
- *Business Survival*: Providing timely evidence concerning safety certificates, financial assets transferred, and other mission-critical documentation can make the difference between continuing in business and being brought to a halt by regulatory oversight.

Document management costs are some of the more difficult company expenses to quantify, but they are very real and very expensive. Add together the financial considerations of the data from the diagrams above and it is easy to justify a hard look at scanning and electronic document management options. And this does not take into consideration the cost benefits that are derived from eliminating manual drawing revisions.

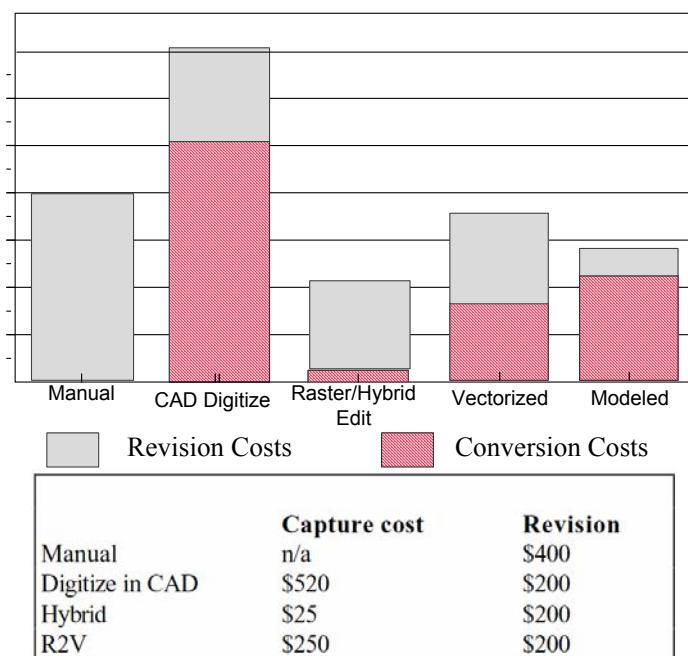
Creating an open environment for information archives requires a one-time cost for scanning paper drawings into an electronic environment. Once implemented, savings are realized throughout all phases of the product life cycle.

A simple cost-benefit example in which a company has 100 drawings with 20 Engineering Change Orders (ECOs) to perform each month can be used to illustrate raster-enabling cost savings. Various labor rates are used for each discipline. Actual numbers should be determined for individual organizations.

Table 6-1 – Cost-Benefit Comparison of Raster-Enabled Process

Task	Times Per Month	Manual Minutes	Raster Enabled Minutes	Burdened Rate	Savings
Find a Drawing	100	1	.05	\$50	\$4,750
Find Related ECO	20	1	.05	\$50	\$950
Approve ECO	20	4	.5	\$50	\$3,500
Update to Rev B	20	3	.1	\$50	\$2,900
Confirm Changes	20	1	.1	\$50	\$900
Distribute Latest Revision	20	3	.1	\$25	\$1,450
Monthly Savings					\$14,450

Figure 6-2 – Cost Analysis of Various Revision Methods



With reduced labor costs and improved usage of CAD, the benefits of revising drawings electronically are clear. What may not be clear is the trade-off between investing in the upfront vectorizing to raw vector CAD, converting to intelligent variational models, or taking advantage of a lower cost hybrid raster CAD system.

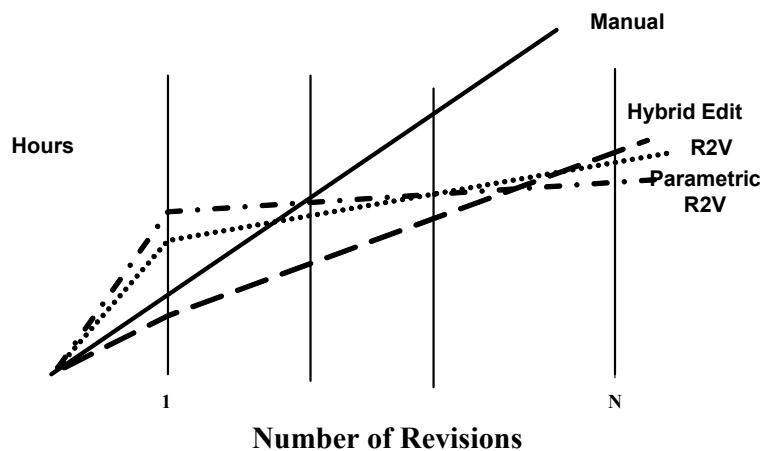
The costs associated with revising drawings are dependent on the method and solution used. The methods evaluated here include manual, CAD digitize, hybrid raster CAD, full vectorization, or conversion to intelligent models. Costs are incurred with each individual revision and include the cost associated with capturing the document to a digital form unless the manual method was used. Therefore, the true cost is calculated by combining the labor rate and the projected time spent on each revision plus the digital transformation expense.

The key driver for determining the cost benefit is the labor savings realized for each method. Any drawing will realize a unique labor savings based on the drawing's quality and the revision method used.

In an article first appearing in *Document Management* magazine, the inherent cost to recreate and revise a complex drawing, using each of these methods were compared. This comparison considered initial capture time, various labor rates, and the time associated with making revisions to the drawing once it had been captured. The results are shown above. The hybrid Raster/CAD approach, which eliminates the redraw, clean up, and verification process, offers the greatest immediate cost benefit for the first revision.

Most drawings realize frequent revisions throughout their active life. A look at a cost benefit analysis throughout that life can also be identified by Figure 6-4. The break-even points for the four primary methods can be determined. A break even analysis can help you understand what method is best for you. The first point is when the hybrid raster/CAD approach becomes beneficial over manual methods (n_2) and is usually attained with the first revision. This is due to the tremendous savings attained with hybrid raster/CAD systems as it allows the modification of old drawings as if they have CAD intelligence. These raster/CAD systems can move, copy, or add selected raster entities fast and accurately with raster snapping features.

Figure 6-3 – Lifetime Cost-Benefit Analysis



The next break even is attained when conversion to full CAD begins its cost benefits over the manual process. This point is highly dependent on image quality, operator training, and tools used. It can be reached as soon as the first revision.

A final break even is realized between the hybrid raster CAD approach and converting to full vector CAD. This is where selective conversion tools have a significant place allowing companies to incrementally convert to CAD as changes become necessary. Cost justification for the various technologies described here is most measurable when based on labor savings in the revision process and improved information access and management.

An estimate of the payback period can be found by dividing the net investment in capture equipment and costs by the annual savings. The return on investment is the annual savings divided by the net investment. An example could be if a company has 500 active drawings and makes an average of 2 revisions (taking 3.5 hours to complete each) per year per drawing. The drafting burden rate is \$35/hour. That adds up to $(500 \times 3.5 \text{ hrs} \times \$35/\text{hr} \times 2)$ \$122,500 per year on drawing revisions.

We can examine this same situation using the hybrid raster editing approach without conversion. Handling, scan time, and enhancement will take about 10 minutes, the burdened operator charge is \$70 including equipment amortization resulting in capture costs of $(500 \times .1667 \text{ hrs} \times \$70/\text{hr})$ \$5,834. There is no vectorization or clean-up time and revisions take 30 minutes per drawing. Revision costs associated with this approach are $(500 \times .5 \text{ hrs} \times \$70/\text{hr} \times 2 \text{ cycles} = \$35,000)$. This is calculated using \$70/hour including the hardware & software costs for a fully burdened CAD operator. This adds up to a potential savings of \$81,666 per year. This example can also be calculated using the vectorization approach. Scan time and burden costs remain the same (\$5,834). Typical conversion time (capture costs) for a large portion of the drawing would take approximately 15 minutes for initial vectorization and 1.5 hours for post processing and QA/clean-up, resulting in a cost of $(500 \times 1.75 \text{ hrs} @ \$70/\text{hr} = \$61,250)$. Revision time is reduced to half an hour as CAD functionality would mainly be used $(500 \times .5 \text{ hrs} \times \$70/\text{hr} \times 2 \text{ cycles} = \$35,000)$. This adds up to \$102,084 – a potential savings of \$20,416 after two revisions. Now that the drawings are fully in CAD, all future revisions would add up to greater savings.

If the total system costs \$30,000 (scanner and software) per year, these examples have a technical payback time of just a few months and the investment is easily justifiable. Keep in mind that although the payback and ROI figures will vary, the cost of manual revisions are already being realized and will only increase.

Companies already spend 7-10% of their expenditures on manual document management processes. A business reinvestment strategy involving the technology presented here can help reduce the incurred costs of managing, revising, and distributing information.

6.1.2 Indirect Benefits

The direct benefits of converting paper to electronic formats are based on labor savings in search and retrieval, revision, and storage costs. However, there are many intangible benefits which include:

- Where contracts are awarded based on accurate and detailed proposals including cost and time estimates, a paper-enabled and intelligent modeling system at the front end yields faster design times and more accurate bids.

- CADD systems increase in value because they are used less for tedious redraw and more for productive design and analysis.
- The enterprise can take advantage of a common electronic database.
- Retrieval and print times are greatly reduced for documents that need merely to be reproduced.
- Information flow can be improved with workflow and e-mail tools.
- The enterprise improves conformance with ISO 9000 or OSHA regulations by instituting better document control procedures.
- Paper drawings increase in value through integration with CADD and EDMS tools.
- Fewer documents are lost, damaged, or misfiled.
- More accurate information becomes available immediately.
- The change process is greatly streamlined.
- Overall work increases in quality.

6.1.3 Purchasing Considerations

6.1.3.1 Scanning Costs

Storage costs are very often less than ten percent of the cost of document conversion. This estimate will provide a basis for assigning the best relative weights and values to the cost of storage and scanning in evaluating system feasibility. For estimating purposes, the commercial scanning cost for backfile conversion (more than 1 million pages) is approximately 5 US cents per page.

6.1.3.2 Software vs. Solutions vs. Process Outsourcing

The value of document capture lies in the process and repository applications that are populated with the captured information. This means that partners and solutions channels take on a particular importance. Suppliers for document capture split into five major types:

Software Vendors

Many document and content capture vendors sell directly to end-users, as well as through a combination of delivery channels. When selling direct, software vendors offer professional services and take responsibility for the final system, including associated hardware, software, and support. Users go to vendors directly with complex requirements that are essentially capture-centric—forms' processing for censuses is a classic example.

General Integrators

General integrators focus on very large project deployments. Usually, a general integrator leads and primes major government and/or international contracts, bringing in specialists to carry out the capture and recognition phases of document conversion.

Specialist Integrators/Solutions Providers

Specialist integrators operate in specific applications, vertical market areas, or geographic regions. They seek to deliver whole solutions into their core market. Such solutions tend to have document capture as an important, but not overriding, component within the overall solutions offered. Examples include accounts payable applications, or insurance claims processing.

Value Added Resellers

Value-added resellers operate in environments where the user is in control of the overall application and is seeking to add document and/or content capture functionality to its facility. Examples might include inventory control, where goods received notes need to be captured and processed, or Human Resources, where Curricula Vitae need to be captured and stored for reference.

Specialist Outsourcers and Application Service Providers

Outsourcing has long been known to be a cost-effective way of converting large volumes of paper documents. Both the public and private sectors have been paying outsourcers, also known as service bureaus, to handle the labor, equipment costs, and technical issues associated with scanning paper. Organizations usually outsource to reduce their operating costs-enabling them to focus their resources on core business processes. The first can be quantified with the proper diligence, while the latter is a strategic decision often taken for soft dollar benefits. There are three main levels of outsourcers and application service providers:

- *Document Capture Bureaus:* These offer functional services, such as scanning, indexing, recognition, and other function-based services. Their output is a document and/or data stream, often on CD and mostly in the archive market. They sell on price and on speed/reliability.
- *Managed Services Providers (MSP):* MSPs set out to deliver documents, data, and process into some part(s) of a business process. As an example, they may take incoming mail, process it, validate it, return anything that needs to go back, handle some customer interactions, and feed the relevant information into their client's business processes. Other examples include document hosting and electronic bill presentment services. MSPs work best where there is an opportunity to implement an annuity or "pay per click" pricing model. The MSP business model is to find replicable services solutions that drive down the cost of service through economies of scale.
- *Business Process Outsourcers (BPO):* The key difference between BPO providers and MSPs is defined by business outcome versus technical or functional output. For example, sending out an invoice is a functional output. Handling the accounts receivable process on behalf of a client is a business outcome. The systems approach is largely the same: the

difference lies in the value of the outcome to the client, and to the level of understanding the BPO provider has of its clients' business. Emerging areas for this level of service are to be found in insurance claims management—with the provider being compensated on reduced costs of processing, rather than numbers of claims processed—and mortgage processing, where the business metric might be based around cost of processing, speed of offer, and/or levels of default.

6.1.3.3 Quality Control

Quality control (QC) systems and processes must be considered, whether or not scanning is outsourced. The use of test targets (as explained in ANSI/AIIM TR34-1996, *Sampling Procedures for Inspection by Attributes of Images in Electronic Image Management (EIM) and Micrographics Systems*) is always a good sign, and records of QC tests are always a good thing to examine. Some outsourcers will QC as they scan, but most will QC post scanning, often during the indexing step. Of course the best QC is through 100% inspection of all images, and even comparing it to the actual document. However, 100% inspection is not often cost justified. A good sampling program can be developed using standards based on acceptable quality levels or random sampling. But the QC process should not end with the outsourcer. Images and indexes should be inspected in a timely manner and there should be a delay before "sign off" on final acceptance.

Other technical components include the outsourcer's ability to transmit the images or return them on CD-ROM or via other transport methods. The technical staff must have the necessary experience: one good sign is if the staff includes one or more Certified Document Imaging Architects (CDIA). Be sure to check out the facility's security and fire protection systems. You should also consider the outsourcer's financial stability, length of time in business, and references. (Keep in mind that no vendor will ever offer a bad reference.) Finally, you should evaluate the outsourcer's other services that might be of benefit, such as document storage, ASP services, or even CM systems sales and integration.

Outsourcing has long proven to be a cost-effective method of leveraging the best in technology and labor utilization. The trick is to go into it with your eyes wide open and do the proper evaluation. This can be done with the help of outsourcing or hardware/software vendors who will tell you what their products or services can do. Many companies exploring outsourcing also use consultants to assist with the analysis and procurement. The right consultants can provide an unbiased view, acting as the organization's advocate. They also bring experience and the tools and methodologies to aid in developing a thorough business case. With a sound business case and outsourcing where appropriate, an organization's capital and human resources can be applied where they will do the most good.

6.2 Applicable Documents

The following specifications, standards, and guides should be reviewed to the extent they apply. Unless otherwise specified, the issues of these documents are those listed in the latest issue of the Department of Defense Index of Specifications and Standards (DoDISS) and supplement thereto. (Unless otherwise indicated, copies of the above specifications, standards, and guides are available from the Standardization Document Order Desk, 700 Robbins Avenue, Building 4D, Philadelphia, PA 19111-5094.)

Department of Defense Standards

- MIL-PRF-28000-A, Digital Representation for Communication of Product Data; IGES Application Subsets & Application Protocols
- MIL-PRF-28001, Markup Requirements and Generic Style Specifications for Electronic Printed Output and Exchange of Text
- MIL-M-28002-B, Raster Graphics Representations in Binary Format, Requirements for Raster Graphics Representation
- MIL-M-28003-A, Digital Representation for Communication of Illustration Data: Computer Graphics Metafile Application Profile
- MIL-STD-2549, Configuration Management Data Interface

Department of Defense Guides

- MIL-HDBK 61, Configuration Management Guidance
- Data Management Guide

Other Government Documents and Publications

The following other Government documents and publications form a part of this document to the extent specified herein.

- Tri-Service CADD/GIS Technology Center (TSTC) (available from Tri-Service Standards DoD Engineer Waterways Experiment Station, 39009 Halls Ferry Rd., Vicksburg, MS 39180-6199)
- AMC Pamphlet 70-25, "Templates for Streamlining Acquisitions"
- Computer-Aided Acquisition and Life Cycle Support (CALS) Specification and Standards Documents
- Federal Acquisition Regulation (FAR)

Non-Government Publications

- ANSI (American National Standards Organization)
- ASME (American Society of Mechanical Engineers)
- IEEE (Institute of Electrical and Electronic Engineers)
- IPO (IGES/PDES Organization)
- IGES (International Graphics Exchange Specification)
- ISO (International Standards Organization)

- STEP (Standard for the Exchange of Product Data)

6.3 EDMS Use Within DoD Organizations

DoD services are increasingly understanding and embracing the need to convert legacy data to accessible electronic formats. Recent Intergraph projects for DoD customers in the document conversion and EDMS fields include:

- Army (AMCOM):
 - Converting AVDS-1790, M113, M1 Abrams, Howitzer, MQM-107 UAV Drone and HMMWV weapon systems drawings for TACOM.
 - Converted Kuwaiti infrastructure drawings and maps for National Ground Intelligence Center (NGIC)
 - Developed Blackhawk Integrated Development Environment (IDE) and converted Blackhawk helicopter technical information files and drawings, including the development of CATIA models. Implemented as part of NAVAIR's JATDI project
 - Conducted functional analyses to determine and optimize business processes using converted data (*e.g.*, Engineering Design, Reprocurrency, and Manufacturing)
- Navy:
 - LPD-17 Integrated Product Data Environment
 - Converted SH60, P3, EA6B, and EOSS weapon systems drawings to vector format from paper/aperture card
 - Converted USN Destroyer Class Ship manuals
 - Drawing management process study for North Island
 - Document conversion analysis study
 - Performing technology assessment for wiring diagrams
- Air Force:
 - Converted C17 and U-2 weapon systems drawings
- Marines:
 - Converting AAV weapon system drawings
 - On-demand data conversion system

6.4 Future Improvement

Technology Advancements Continue

Backfile conversion has benefited from a number of recent advances in technology and the progress in internationally accepted standards and codes of practice:

- Recent developments in Intelligent Document Recognition IDR techniques provide more cost effective approach to capturing the metadata and automating the conversion process.
- Storage costs have plummeted for both RAID and optical discs and their performance improved. Newer storage concepts, such as NAS and SAN, are also falling in price. These are in addition to Direct Attached Storage, which is in common use. They provide improved access and performance to the archive with built-in security and backup. Manufacturers are also addressing the need for backward compatibility of optical disks. For example, a new Hewlett Packard 5¼-inch optical drive has a capacity of 9.1 gigabytes but still reads the original 650 megabyte disks, launched in 1988. Manufacturers such as Hewlett Packard have committed to providing similar backward compatibility with a new range of 40 gigabyte optical disks.
- XML and JPEG 2000 are examples of standards that are future proofing against changes in operating systems and line of business operations. All mainstream suppliers of technology and software are incorporating these standards.
- National and International codes of practice to address the legal admissibility of electronic documents are now in place nationally in most European countries and an international standard is in draft format. EDMS and RMS suppliers are ensuring their systems can meet these requirements.

Document Capture is a Tactical Business Advantage

As document and content capture mature and become safer to buy, so the nature of the decision-making process for its implementation must change. Systems are increasingly justified on straightforward business metrics of cost-saving, improved performance, and better customer service. In many sectors, such as insurance claims processing and regulatory compliance, it has become difficult to implement new solutions without using document and content capture techniques and technologies.

A Mere Archive Is Insufficient

Research shows that more paper is scanned for use by a business process than for archiving alone. Between 1998 and 2000, the proportion of applications installed for archiving and repository purposes fell from 62% of the overall market to just 40%. The developing value proposition for metadata-enabled content capture and document retrieval lies in the way that it allows organizations to search for legacy data many times faster than before, with high integrity and accuracy, and in a way that complements the business process being served.

Implement ODR Today—But Plan for IDR Tomorrow

Optical Document Recognition actually works for “conventional” forms processing. ODR is deployed in tens of thousands of installations in applications ranging from census projects through healthcare claims processing to market research and government forms, among others. As performance continues to improve and as hardware prices in particular continue to fall, the range of potential applications continues to grow.

As semi-structured document processing develops, however, the range of applications that can benefit from automated data extraction from documents will grow incrementally. Today, IDR is achieving success in the invoice processing market. Soon, new applications in document processing for knowledge management will come to dominate the market for document capture.

Automatic Classification by Content and Context

Automatic classification is a major driver of document capture, enabling increasingly automated indexing and consequent massive cost reductions. Automatic classification systems, merged with an organization’s workflow to capture everything the organization produces, go beyond just analyzing content to take into account:

- Context mapping
- Subject, thesauri, etc.
- Implicit and explicit profiling (personalization)
- Relationship mapping — between documents, documents and processes, documents and roles, and documents and people

7 Conclusion

The Corps of Engineers' vast library of sometimes irreplaceable information is difficult to use. Site visits reveal that Corps archives are deteriorating and poorly indexed, and that documents are too frequently time-consuming or even impossible to find. Hard-copy document archives drain the Corps of large amounts yearly in real estate costs, overhead costs, and work delays. The Corps' collected knowledge of decades is disappearing at an unknown rate, and lost information must be laboriously recreated, imposing additional costs in data gathering and labor hours – if the information can be recreated at all.

Within the Corps of Engineers today, many departments are scanning and converting their legacy hard-copy documents to digital formats. Some knowledgeable departments are using informed processes to establish sound conversion strategies. But most departments are neither collecting nor planning to collect metadata, the attributes that describe the electronic output and make it most valuable. Electronic documents having no searchable attributes lose most of their value because they are at least as hard to find as their hard-copy equivalents. To recapture the value of converted documents, the Corps must focus on a concerted, uniform effort toward setting standards in document capture, metadata collection, document and metadata storage, metadata search, and document retrieval.

Conversion standards must take into account input document types, needed metadata attributes, standardized output types, and community-wide access. Standards must be evaluated to ensure all relevant data is captured from the input document; aerial photographs, for example, have resolutions an order of magnitude greater than the common 300 dpi. Further standards must be implemented that categorize the output documents, based on document type, department, discipline, or some relevant bit of data by which to sort the information.

The most important step in ensuring an electronic document archive's lasting value is standardizing and collecting the documents' metadata. A highly useful set of metadata attributes is that set required by ARIMS; if collected, this metadata can be directly transferred into ARIMS upon adoption. Additional attributes are also worth considering; each document type may contain useful attributes useful to the Corps that are not required by ARIMS.

The Corps must define additional standards for storing, searching for, and retrieving the archived documents with their associated metadata. Sharing the electronic archive will continually add worth to the stored documents as the valuable information contained in them is reused instead of replicated. Increasing accessibility to existing information has hidden workflow benefits that can be increased with a workflow management system. Other benefits can be increased by implementing document and/or records management.

Technology already available off the shelf has capabilities the Corps can use to meet these goals. Though funding is generally not available for overhead functions such as standalone document conversion tasks, dollars may be easier to obtain if improving a workflow for the delivery of a project provides document conversion as a byproduct. Purchased and implemented in a planned and consistent manner, conversion of technical data from paper to digital formats can make document and information reuse within the Corps much more efficient.

The goals of reduced storage costs, greater productivity, shorter design and drafting time, shorter information search and retrieval time, less repetition of effort, greater consistency with past work, and preservation of valuable information are all achievable. The cost reduction realized from achieving these goals stands to significantly outweigh the cost of implementation.

Appendix A. Glossary

Term	Meaning
AEA	Army Electronic Archive.
Analog	A method to represent information by continuously varying signals such as an electrical current in the shape of a wave.
ANSI	An acronym for American National Standards Institute. The institute is a private organization that develops, maintains, and publishes industry standards in the United States.
Aperture Card	A card that has a 35 mm microfilm frame attached to it in a cut-out area of the card. Aperture cards are typically used for archival storage of printed or other hard copy materials, as the storage life of archivally prepared film is longer than the usable life of many original materials, such as paper drawings.
Aperture Card Scanner	An optical scanning device that converts aperture card film images into pixels.
API	An acronym for Application Programming Interface.
Archive	Copy of the data and indexes of a database onto a new medium, usually a tape of a different physical device from the one that stores the database.
ARIMS	Army Records Information Management System, the system expected to replace MARKS (q.v.).
ASCII	An acronym for the American National Standard Code for Information Interchange. ASCII is a code that uses seven bits to represent graphic and control characters. In general, an ASCII file is a text file.
Attribute	A piece of descriptive information about an electronic file: its author, its content, etc. The set of attributes is referred to collectively as metadata.
Backup	To make a duplicate of a computer file on another device or on tape in order to preserve existing work in case of a computer failure or other mishap.
BCR	Bar Code Recognition, used to translate "thick-thin" vertical bar patterns into a numerical sequence, usually for inventory control or point-of-sale purposes.
Binary Code	A code that represents information as a sequential series of 0s and 1s.
Bit	A unit of binary data, either a 0 or 1, represented by the presence or absence of an electric signal.
Bitmap or Bit Map	2D array of pixel elements.
Bulk Loading	The process of transferring database information from one system to another.
Byte	Group of 8 adjacent binary digits (bits) that represent a character.
CAD	Computer-Aided Design.
CADD	Computer-Aided Drafting and Design.
CALS	Computer-Aided Acquisition and Logistics Support initiative, a Department of Defense (DoD) program setting standards for electronic interchange of information. An initial focus of CALS is publishing standards. Beginning in 1990, all weapon system documentation must be submitted to the DoD in electronic form.
Cartographic	Pertaining to the technique of making maps or charts.
CCD	Charge-coupled device, a semiconductor that can collect, store, and move charges in packets. CCD scanners reflect light off the document onto a linear array of photosensitive cells that convert the reflected light into a binary bit stream. In this way, the document is converted into series of thin slices which in turn are divided into small pieces. The thickness of the slice and the number of individual divisions of the slice determine the scan resolution. When reassembled by a computer, the document takes on the appearance of a large grid (raster image). Each grid can have a value of on or off (black or white) value for brightness (grayscale) or even a color value associated with each element within the grid (pixel.)

CCITT	An acronym for Consultative Committee on International Telephone & Telegraph. CCITT is a committee formed in 1938 that sets the international standards for the hardware and communications protocols for data and voice transmissions.
CCITT	Consulting Committee for International Telephony and Telegraphy, an organization that develops international communication standards. To date, four such standards have been designated: Group 1: A bit-map raster standard for use in facsimile transmission of analog signals. Group 2: Huffman run-length code is a one-dimensional means to compress raster files for use in facsimile transmission of analog signals. Instead of storing each pixel, as in Group 1, the Huffman run-length code designates the coding and decoding of a binary transition, such as a change from a black pixel to a white pixel. Group 3: Relative Element Addressing (READ) code is a 2D means of compressing 200-dots-per-inch raster files for use in facsimile transmission of documents. The READ code further enhances the Huffman run-length code by collecting statistics on the length of runs and determining the number of most frequent occurrences. This redundancy reduction technique is possible by digital processing of black- and-white documents. Group 4: This standard for digital, high-speed data transmission is just emerging. There no longer will be requirements for digital-to-analog or analog-to-digital conversion. The speed of transmission will increase and the alternative transmission means, such as satellite transmission, will become available.
CD-ROM	Compact Disc Read Only Memory, a type of optical disk storage.
CEEDMS	Corps of Engineers Electronic Document Management System.
CEERIS	Corps of Engineers Electronic Recordkeeping Information System.
Centerlining	The process of enclosing each raster line with a linear polygon (rectangular fenced area) and then calculating the centerline of that polygon.
Client	A software process that makes use of the services of another software process.
Compression	The process of reducing the amount of storage required to represent any array or collection of data.
Concurrency	The ability of two or more processes to access the same database simultaneously.
Continuous Tone	Raster data where shades of gray or brightness are associated with each pixel. Each pixel is represented by multiple bits.
Conversion	The process of changing a physical document or analog equivalent to electronic form, such as raster or vector data.
Converted Document	The output of a conversion process.
Data Format	Determines how pixels are stored. Typical formats are uncompressed bitmap (1 bit per pixel), run-length encoded (RLE), CCITT Group 3 and Group 4. See CCITT for more examples.
Data Integrity	The process of ensuring that data corruption does not occur when multiple users simultaneously try to alter the same data. Locking and transaction processing are used to control data integrity.
Database	A collection of information (contained in tables) having predetermined structure and organization that can then be communicated, interpreted, or processed by a specific program.
Database Application	A program that applies database management techniques to implement specific data manipulation and reporting tasks.
DBMS	An acronym for Database Management System. DBMS is all the components necessary to create and maintain a database, including the application development tools and the database engine.
Decompression	The process of expanding compressed raster data.
Digital	A method to represent information by discrete or individually distinct signals, such as bits.

Digitize	<p>The act of converting information on drawings or documents to vector-based electronic form. Digitizing methods include:</p> <ul style="list-style-type: none"> ▪ Manual—the process of digitizing a drawing using a digitizing tablet. While this is faster than redrawing totally, with many CAD systems it is still labor-intensive and error-prone because the operator is constantly moving from the tablet to the screen. Once a drawing has been digitized into the CAD system, the operator must check the original drawing against the new drawing to verify each component has been included. ▪ Heads-up—the process of digitizing a drawing by tracing over a scanned image on the screen. This is less expensive and faster since one scanner can support 10, 20, or even 100 draftsmen, and the time-consuming task of drawing validation is eliminated. With this method, the draftsman, upon completion of a job, instructs the software to display the original information (raster) in one color (e.g., green) and the new data (vector) in another color (e.g., red). If the red disappears, then the drawing is complete and accurate. ▪ Automated—the process of using a computer program to convert a scanned image of a drawing to a vector format. The method is dependent on algorithms that recognize patterns of adjacent pixels and determine the vector-based elements (lines, arcs, text) that they probably represent. The method is reasonably successful on documents that are clean (no dots or speckles) and where lines do not cross each other, as in topographic contour maps. ▪ Semi-automated—similar to the category of “Automated” above, except that the operator is prompted to make a decision where the program cannot determine the direction that a vector element should proceed. This occurs, for example, when lines cross or for lines represented as dashed or dotted. This method is the most efficient method for digitizing most types of drawings.
Directory	A logical construct for organizing information stored on a disk; disk directories can contain files and other directories.
Disk Mirroring	Storing the same data on two disks simultaneously. If one disk crashes, the data is still usable on the remaining disk.
Disposition	The eventual fate of an asset, including a document.
Dithering	The process of simulating continuous-tone data in black and white using pixel patterns of various sizes.
DMS	An acronym for Document Management System.
Document	Any physical or electronic free-standing unit of information, such as a report, drawing, spreadsheet, memo, or other information represented either on paper, microform, or electronic media. Electronic documents include information types such as sound and video files, but do not include record-based information such as a single row or element from a database.
Document Management	A system of storage and recordkeeping by which electronic documents are “filed” and kept. Provides a means of searching for and retrieving documents.
DKE	Distributed Knowledge Environment.
DPI	Dots per inch, a way of measuring the resolution of a scanning input device, a display device, or an output printer/plotter device.
EBS	Electronic Bid Sets.
EDMS	Electronic Document Management System.
Ethernet	A linear bus topology network originated by Xerox Corporation which uses CSMA/CD access protocol and allows multiple protocols such as XNS and TCP/IP to coexist on the same transmission system.
Facsimile (FAX)	Equipment that transmits and receives hard copy over normal telephone lines.
Fast Recovery	Automatically restoring a database to its state at the end of the last completed transaction after a non-destructive failure of the computer system (for example, a power failure but, not a disk failure).
Fault Tolerance	The ability of a system to resist failure and loss of data. It includes features such as fast recovery and disk mirroring.

Feed time	The time required to move or feed a document through the transport mechanism of a roll-feed scanner.
File	A collection of related information that is stored on a disk. A file on a disk has a name through which it is accessible. Related files may be grouped together in a common directory.
File Server	A computer running a specialized program that provides network users with access to shared disks or other mass storage devices. Through access controls, a file server facilitates controlled access to common files and applications.
Flatbed	A mechanism for holding the original document on a flat surface. In a flatbed scanner, the document is stationary and the light source and cameras move.
FTR	An acronym for Full Text Retrieval.
Grayscale	A way of representing or printing continuous-tone data.
GUI	An acronym for Graphical User Interface.
Hybrid Raster/Vector	A file containing both vector and raster data.
ICR	Intelligent Character Recognition, a means of recognizing printed material and converting it into word-processor text.
IDR	Intelligent Document Recognition, a means to automatically identify a document's type and its related attributes from the layout and structure of the document.
Image Enhancement	The processing of an image so that the result is more visually apparent, accurate, or suitable than the original image.
Incremental Archiving	Archival of only those parts of the data that have changed since the last archive.
Index	A file containing pointers to rows of data. Indexes can speed ordering of rows and optimize the performance of database queries.
Internet	The public Internet, an unregulated, interconnected system of computers that communicate information to one another worldwide.
Intranet	An interconnected system of computers restricted to a single enterprise.
ISO	An acronym for International Standards Organization.
Knowledge Management	A system of organizational processes that seek synergistic combination of data and information processing capacity of information technologies, and the creative and innovative capacity of human beings. A means of cataloging, preserving, and making accessible the collected knowledge assets of an organization.
LAN	An acronym for Local Area Network. Local Area Network shall mean a group of Stations and other business computing equipment in a single location or geographical area and connected by a data communications system allowing them to transfer electronic data between them.
Legacy	Outdated, outmoded, discontinued, or being phased out.
LPI	Lines per inch. See dpi (dots per inch).
Metadata	A collection of information about information, or a list of characteristics summarizing an information-containing package.
MICR	Magnetic Ink Character Recognition, a method by which a machine recognizes (usually) numerals printed with ink having magnetic properties.
Mil	One thousandth of an inch.
Mylar®	Mylar® is an extraordinarily strong polyester film that grew out of the development of Dacron® in the early 1950s. During the 1960s cellophane gave way steadily to Mylar with its superior strength, heat resistance, and excellent insulating properties. The unique qualities of Mylar made new consumer markets in magnetic audio and video tape, capacitor dielectrics, packaging and batteries possible. By the 1970s, Mylar had become DuPont's best-selling film, despite mounting competition. Some varieties of Mylar are suitable for printing and all are much more durable than paper.
NARA	National Archives and Records Administration, which oversees the management of federal government records, including presidential diaries, historic correspondence, and a display of presidential gifts from around the world.

OCR	Optical character recognition, the technique by which a machine recognizes printed or photographically recorded characters using a combination of image capture and electronic logic, then converting the characters to binary digital codes for storage, transmission, etc.
ODR	Optical Document Recognition, a process of by recognizing and extracting data from paper forms in order to feed that data into a business application.
OMR	Optical Mark Recognition, a technique for recognizing pre-defined shapes in pre-determined positions.
OSI	An acronym for Open Systems Interconnection. OSI is a model developed by the International Standards Organization to describe a network that would be open to equipment from many vendors.
Pel	Same as pixel, a means to describe the raster density in terms of black-and- white characters or images, such as pel/cm or pel/in. Example: a <u>s</u> can density of 200 lines per inch means each pel represents an area of .005 × .005 inches. If the image is black or white, there is 1 bit per pel. If the image is a halftone, there could be 6 bits. If the image is color, the number of bits per pel would increase depending upon the color and color gradation.
Pixel	Picture (X) Element, single element of digital data
RAID	Redundant Array of Inexpensive Disks. The term is used to describe a disk/data management system utilized on multi-processor machines. RAID has different capabilities including 0,1,3 and 5. Each capability indicates a function of RAID technology including data striping, data mirroring and hot disk replacement.
Raster	A means of representing an image by an array of pixels.
Rasterization	The process of converting text or images to pixels.
RDBMS	An acronym for Relational Database Management System. RDBMS is the software providing user capability for organizing, storing, and manipulating data in a database.
Redlining	The process of marking a drawing to show changes.
Relational Database	A model that uses columns and rows on tables to establish relationships between data. Relational databases have at least one table.
Resolution	A way to describe the readability of a printed sheet or drawing for scanning or plotting. See dpi.
Revision	A document iteration that is maintained separately and simultaneously with other iterations of the same document, and which is directly accessible to the user of the EDMS. EDMS document uniqueness is based on the combination of document name and revision.
RHA	Records Holding Area.
RMS	Records Management System. An RMS generally holds immutable records information.
Run-Length Encoded (RLE)	Compressed raster file format. One typical application is for use in encoding data for facsimile transmission of analog signals.
Scan Document	There are two standards for scan-size documents, one defined by ANSI for American-size documents and drawings and the other by ECC for European-size documents and drawings.
Scan Time	Total time to convert text or graphical information to electronic raster form.
Scanner	A device that translates the spot densities of a document into pixels that form digitally encoded raster images.
Scanner	An optical device that converts hard copy images (paper, film, or aperture) into pixels
Server	A computing system connected to a LAN or WAN, and whose functions are generally shared by multiple users indirectly from other computing systems.
SGML	Standard Generalized Markup Language. An enabling technology used in applications such as HTML.
SQL	An acronym for Structured Query Language. SQL is an English-like query language used to access a database.

Supported	Utilizes compatible protocols, operating systems, software, device drivers, etc.
TCP/IP	An acronym for Transmission Control Protocol/Interconnect Protocol.
Thinning	In OCR, a stylized character pattern in computer code form that is matched against scanned pixel data to recognize a letter or number. The process of vectorization by which groups of pixels are reduced until they are one-pixel deep.
Thresholding	The process by which pixels are saved as on or off when scanned. With I/SCAN, the boundary determines at which point Threshold differentiates between these two functions. For example, if a user designates 120 as the boundary, all data lighter than 120 becomes white and all data darker than 120 becomes black.
Throughput	The number of images of a specified size or type that can be scanned in a given amount of time.
TIFF	Tagged Image File Format, a series of file formats for saving, editing, and printing scanned images.
Transaction	A collection of one or more SQL statements that is treated as a single unit of work. If one of the statements in a transaction fails, the entire transaction can be canceled. If the transaction is successful, the work is accepted.
UNIX	A computer operating system developed by AT&T Bell Laboratories.
Vector	A method of describing an image as a series of points and a description of how to display the data represented by the points. Examples of vector information include lines, arcs, linestrings, b-splines, and text (glyph shapes and positions).
Vectorization	Conversion of pixels to vector data, generally in one of two ways: centerline or edge.
Version	A document iteration generated by the EDMS for backup purposes, and which is not directly accessible by the end user. In the event of system or operator error, previously retained versions can be restored administratively to become available for direct user access.
View-Only	A software capability that allows a user to view but not modify or annotate an electronic document.
WAN	An acronym for "Wide Area Network." WANs generally consist of 2 or more local area networks (LANs) connected together by telecommunications bridges and/or routers, as well as fiber optic connections.
Workflow Instance	A working copy of the workflow template.
Workflow Management	The automation of business procedures or "workflows" during which documents, information, or tasks are passed from one participant to another in a way that is governed by rules or procedures.
Workflow Template	The master copy of a workflow.
WORM	Write Once Read Many, generally applied to optical media.
XML	Extensible Markup Language, a simple, very flexible text format derived from SGML. Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere.

Appendix B. Metadata

Metadata, commonly referred to as “attributes,” is descriptive data about other data, or in this case, documents. Attributes provide an ability to better define of the scope and relationships between documents and provide the infrastructure to enable efficient searches for and retrieval of documents. As a result, the selection of attributes, and the determination of which ones are “mandatory” and which ones can be “optional,” is extremely important in the definition of an EDMS. The following table lists ARIMS attribute requirements, recommended for Corps consideration:

Table B-1 – Indexing Data Elements

Data Element Name	Data Type	Req?	Size	Comment
DOC_AUTHOR	Alpha	Y	50	System generated, editable, pick list.
DOC_CONTRACT_NO	Alpha-Numeric	N	25	System-generated default with pick list.
DOC_DATE	Date	Y	11	4-digit year. System generated, editable.
DOC_FORMAT	Alpha-Numeric	Y	25	Originating application and version. System generated.
DOC_ID	Alpha-Numeric	Y	25	System generated, unique, unalterable.
DOC_LOCATION_ID	Alpha-Numeric	N	5	Location code denoting the offline or near-online storage container containing the document. Does not appear for EDMS documents.
DOC_MARKS_OR_ARIMS_NO	Alpha-Numeric	Y	25	System-generated pick list, editable by authorized personnel only.
DOC_MEDIUM	Alpha-Numeric	Y	20	Defaults to “Electronic” with pick list available for other media. If offline or near-online medium selected, prompt appears for DOC_STORAGE_CONTAINER_ID.
DOC_OFC_SYMBOL	Alpha	Y	20	System generated default with pick list.
DOC_PROJECT_NO	Alpha-Numeric	N	20	System-generated default with pick list.
DOC_REGULATORY_PERMIT_NO	Alpha-Numeric	N	25	System-generated default with pick list.
DOC_STORAGE_CONTAINER_ID	Alpha-Numeric	N	20	Container for offline or near-online document. Does not appear for EDMS documents. Prompt appears for DOC_LOCATION_ID.
DOC_SUBJECT_OR_TITLE	Alpha-Numeric	Y	255	User generated or system captured.
DOC_VITAL_RECORD_INDICATOR	Logical	Y	1	Is document needed to operate during or immediately after an emergency? Does the document protect the rights and interests of the Government or public?

Table B-2 – Non-Indexing and Table Data Elements

Data Element Name	Data Type	Req?	Size	Comment
CONTRACT_NO (K)	Alpha-Numeric	Y	20	An official Corps contract.
CONTRACT_TITLE	Alpha	Y	255	
CONTRACTOR_NAME	Alpha	Y	150	
DOC_TYPE	Alpha	N	25	See Appendix C. Should be a system-generated pick list.
EMPLOYEE_FNAME	Alpha	Y	25	
EMPLOYEE_LNAME	Alpha	Y	25	

Table B-2 – Non-Indexing and Table Data Elements

Data Element Name	Data Type	Req?	Size	Comment
EMPLOYEE_ORGANIZATION_OFC_SYMBOL	Alpha-Numeric	Y	20	
EMPLOYEE_USER_ID (K)	Alpha-Numeric	Y	10	UPASS Administrator maintains this information.
LOCATION_CITY	Alpha	Y	20	Where the document is, physically.
LOCATION_ID (K)	Alpha-Numeric	Y	5	
LOCATION_PO_BOX	Alpha-Numeric	N	10	
LOCATION_STATE	Alpha	Y	2	
LOCATION_STREET_ADDRESS_1	Alpha-Numeric	Y	25	
LOCATION_STREET_ADDRESS_2	Alpha-Numeric	N	25	
LOCATION_ZIP_CODE_1	Alpha-Numeric	Y	10	
LOCATION_ZIP_CODE_2	Alpha-Numeric	N	10	
MARKS_ARIMS_DISPOSITION_CODE (K)	Alpha-Numeric	Y	5	Corps business rules dictate the disposition rules be further refined into codes; <i>e.g.</i> , rule K6 may be assigned a code R2, indicating the office should review the records/files after 2 years to determine whether the files can be destroyed earlier than 6 years. Rules are assigned and updated by the Records Manager.
MARKS_ARIMS_DISPOSITION_RULE	Alpha-Numeric	Y	50	The disposition rule assigned to a MARKS/ARIMS number; <i>e.g.</i> , K6 = Keep in CFA for up to 6 years, KE6 = Keep in CFA for up to 6 years following an event, T10 = Transfer to RHA at any time but destroy after 10 years, etc.
MARKS_ARIMS_NO (K)	Alpha-Numeric	Y	25	
MARKS_ARIMS_TITLE	Alpha	Y	80	
MEDIUM_CODE (K)	Alpha-Numeric	Y	5	E = Electronic N = Near-online (on-site CD, DVD, hard disk, etc.) O = Offline (physically located off-site)
MEDIUM_TYPE	Alpha	Y	20	Explains MEDIUM_CODE.
ORGANIZATION_NAME	Alpha-Numeric	Y	35	For example, Environmental Resources Section.
ORGANIZATION_OFC_SYMBOL (K)	Alpha-Numeric	Y	20	For example, CENWS-PM-PL-ER.
PROJECT_LOCATION	Alpha	Y	50	An authorized Corps project.
PROJECT_NO (K)	Alpha-Numeric	Y	20	
PROJECT_STATE	Alpha	N	2	
PROJECT_TITLE	Alpha	Y	255	
REGULATORY_PERMIT_APPLICANT_NAME	Alpha	Y	50	
REGULATORY_PERMIT_NO (K)	Alpha-Numeric	Y	25	
REGULATORY_PERMIT_WATERWAY_NAME	Alpha-Numeric	Y	35	
STORAGE_CONTAINER_ID (K)	Alpha-Numeric	Y	20	
STORAGE_CONTAINER_TYPE	Alpha-Numeric	Y	20	

Appendix C. Document Types

The general definition of a document is a physical or electronic free-standing form of information. It may be a report, drawing, map, spreadsheet, memo, or other information represented on paper, microfilm, or electronic media.

Details of the definition of a document are important for the implementation of a document management system. A review of several different Districts and Corps documents relating to documents and records has led to the following recommendations:

- For engineering drawings, plans, charts, maps, photographs, and other similar items, a “document” shall be comprised of a single sheet or item. Groups of these types of documents, such as in a set of construction plans, shall be called a “set” which is composed of one or more individual documents.
- For reports, correspondence, spreadsheets, automated presentations (PowerPoint, etc.) and all documents in the general category of office automation, a “document” shall be the set of pages, regardless of their physical size or quantity, that would comprise, for example, an entire “report.”

These definitions of the term “document” lead to the occasional need to have parent and child relationships, such as when a drawing that has a life of its own as a document, due to its existing as a component of a set of construction plans, is included in a report as an attachment. Therefore, document retrieval methods must support these definitions of a document and of parent-child relationships between documents.

The relationships between documents and various forms of accounting or project related numbers can be complex. As an example, it is quite common for a given set of drawings issued for a construction contract to be associated with multiple sources of funding, or for a single funding authorization to be associated with multiple sets of plans or studies. As a necessary result, document management systems must be able to support these types of one-to-many, many-to-one, and many-to-many relationships.

Document management systems must address the difference between a source document in an electronic format and the printed or plotted image. This becomes an important factor as the renditioning process becomes more complex. Of particular interest is the process of plotting a drawing or map from a source CAD file. It is common for a plotted image of a CAD drawing to be composed of portions of several different CAD files, and for the image to be deliberately changed during the plotting process. As a result, there needs to be a distinction between the source CAD files and the plotted or “renditioned” image that may later be issued as a part of a set of construction plans.

File Type refers to the structure of the electronic file that contains the informational content of the document. In many instances, the file type is associated with the particular program that is used to create or edit the content of a particular file type. In the context of the Windows operating system, the three character extension of the file (the three characters to the right of the decimal) are used by the operating system to determine the file type and which program should

be used to open the file when the user double-clicks on it. In some instances, multiple programs could be used to open a given file type. Examples include:

Table C-1 – Document Format Examples

Extension	File Type / Program	Comments
.doc	MS Word	Microsoft Word document
.xls	MS Excel	Microsoft Excel spreadsheet
.pdf	Portable Document Format	Adobe Acrobat file format. This is a common file format used to encapsulate documents of other types. The free Acrobat Viewer is used to view and print, but not modify, the file.
.dgn	MicroStation	A CAD file format used by the MicroStation CAD program from Bentley Systems.
.dwg	AutoCAD	A CAD file format used by the AutoCAD CAD program from Autodesk.
.tif	TIFF raster file	The tiff file format is a standard format (one of many) for raster files. Tiff files can be read and edited by a variety of raster editing programs and many programs use this format as an output format for printing or plotting of other proprietary format files.
.cal	CALS raster file	The CALS file format is a standard format (one of many) for raster files. CALS files are used for document imaging and therefore only store black-and-white, 1-bit image data. The CALS raster file format is defined primarily in the following military standards documents: <ul style="list-style-type: none"> ▪ MIL-STD-1840A, Automated Interchange of Technical Information ▪ MIL-R-28002A, Requirements for Raster Graphics Representation in Binary Format

Some retrieval systems support “external” documents. An external document is an indexed document that is not contained within the system, but rather is stored externally. Examples include a map that exists only as a large, wall-mounted hanging, or a submittal board of an architectural finish. In these cases, the file type would be “external,” rather than an electronic file format.

In the context of document conversion, the document “type” refers to the business purpose or informational content of the document, regardless of the format (File Type) in which the document is stored. Document types are general categories, with a sub-type that further identifies the contents. The reason is that handling of the document depends on a greater definition of the contents than is provided by the general categories of a document type. A sample list of document types and sub-types follows.

Table C-2 – Proposed Document Types

Document Type	Category	Comments
Aerial Photograph	Graphic	Should be scanned at very high resolution.
Appraisal	Real Estate	
Bid	Drawing	
Brief	Legal	
Briefing	Informational	
Budget	Guidance	
Circular	Informational	
Complaint	Legal	
Correspondence	Multiple	
Decision	Legal	

Document Type	Category	Comments
Deed	Legal, Real Estate	
Delivery Order	Contractual	
Directive	Guidance	
Drawing/Sheet	Contractual, Other	
Executive Summary	Informational	
Fact Sheet	Informational	
Form	Informational	
Funding Appeal	Funding	
Funding Request	Funding	
General Order	Guidance	
Graphic	Graphics	Audio visual, computer generated graphic, or other presentation graphic, as differentiated from Drawing/Sheet, Map/Chart, and other document types stored electronically as graphic images.
Historical Document	Informational	
Information Paper	Informational	
Inspection Report	Multiple	
Invoice	Real Estate	
Lease	Real Estate	
Leave	T&A Document	
Letter	Multiple	
Letter of Instruction	Guidance	
License	Real Estate	
Manual	Informational	
Map/Chart		
Memorandum	Contractual	
MIPR	Funding	
Motion/Pleading	Legal	
MOU/MOA/Charter	Contractual	
News Release	Informational	
Opinion	Legal	
Orders	T&A Document	
Pamphlet	Informational	
Permanent Order	Guidance	
Permit	Real Estate	
Photograph	Graphics	
Policy Memorandum	Guidance	
Proposal	Correspondence	
Public Notice	Informational	
Regulation	Guidance	
Report/Study/Plan	Informational, Contractual	
Request For Information	Contractual, Other	
Response	Informational, Contractual	
RFP/RFQ	Contractual	
Solicitation	Contractual	
Specification	Contractual	
Statement Of Work	Contractual	
Survey		
Task Order	Contractual	

Document Type	Category	Comments
Travel Request		
Travel Voucher		

Appendix D. 1999 ADCS Project Synopsis

SUBMITTAL FOR BULK CONVERSION SUPPORT

Program Name: Bulk Conversion of Data for Corps of Engineers Electronic Document Management System (CEEDMS)

Military Service: U. S. Army

Service Priority: 1

Program Sponsor: U. S. Army Corps of Engineers

Name of Command: HQUSACE

PMs/POC Names:

Program Manager: Linda Worthington
(202) 761-0332
Linda.K.Worthington@hq02.usace.army.mil

Technical Manager: Preston Ferguson
(304) 529-5215
Preston.L.Ferguson@lrh01.usace.army.mil

Support Staff Leader: Raymond Rogers
(304) 529-5370
Raymond.R.Rogers.ii@lrh01.usace.army.mil

Categories of Services Supported by the Program (check all that apply):

Conversion:

Low-end conversion (raster / pdf): ☒

R2V conversion: ☒

SGML / ETMs with illustrations: ☒

Simulation Based Acquisitions:

IDE integration to DoD Std. Sys: ☒

Product Data Model: ☒

Data Repositories: ☒

Software Tools: ☒

Standards: ☒

Title

Corps of Engineers Electronic Document Management System (CEEDMS)

Objective

The primary objective for conversion of Corps of Engineers Documents and drawings is to electronically store the files in a Corps of Engineers Electronic Document Management System. CEEDMS will consist of a commercial-off-the-shelf (COTS) electronic document management system configured with Corps standards. The initiative to acquire CEEDMS began on 7 May 1999. In order to take full advantage of the CEEDMS deployment, the Corps will need to convert a large number of 8½ × 11, 8½ × 14, C-sized, and E-sized documents to be entered into the system. The conversion of hard-copy documents and drawings will significantly reduce the cost of doing business throughout the Corps of Engineers, as well as reduce the cost of design duration, file storage, retrieval, and reproduction.

Projected Process Improvements/Benefits

DESIGN – A large quantity of the design work performed within the Corps of Engineers is maintenance or additions to existing structures or property. Frequently this design work utilizes previous design drawings from original or previous construction, and the majority of these drawings are hard-copy mylars or blue-line prints. Significant cost savings will be achieved if these hard-copy documents are converted to electronic documents. Once converted, the electronic documents could be used with new designs and the files could be transferred to contractors electronically.

STORAGE – Throughout the Corps, organizations store millions of documents of various sizes. These documents must be kept on site in most cases to be readily accessible. Every year the Corps pays for the floor space to house these documents. Although a thorough investigation of the total cost of floor space to store documents has not been performed for the entire Corps, Headquarters USACE estimated that they could save about one million dollars per year if they eliminate the need for storing hard copy documents housed at their facility. This figure only considered one facility (Headquarters USACE) with 1000 employees out of sixty-three with 35,000 employees. Reduction of hard copy document storage will significantly reduce the cost of floor space within this organization.

RETRIEVAL – Retrieval of documents for review or reproduction in the existing environment at the Corps can be a tedious and time-consuming task. Often, the location of files is difficult to determine, or access to the documents is limited. Documents in some cases are stored in large file rooms in remote locations within Corps facilities. It may take anywhere from 5 to 30 minutes for a Corps employee to retrieve a document, and that is not considering that many documents may not be stored on site. Some documents may be stored at other locations or Corps offices. This increases the time to acquire documents by hours or even days. The goal for the Corps of Engineers is to implement CEEDMS and scan critical documents and drawings for insertion into the system. By implementing a program such as this, Corps employees will have the ability to access documents anywhere in the organization by utilizing CEEDMS workstation or web client software regardless of their location or the location of the electronic file. Implementation of CEEDMS has the potential to greatly reduce document retrieval costs.

REPRODUCTION – The benefits to converting hard copy documents to electronic format will not only positively affect document storage and retrieval costs, but document reproduction as well. Once documents are converted to raster or vector format and indexed in CEEDMS,

reproduction of these documents will be reduced to a printing or plotting procedure. Again, this will improve employee productivity, thereby reducing costs.

Benefits

- To collect, index, and integrate legacy documents and information into a Distributed Knowledge Environment (DKE).
- To use existing data conversion facilities that limits and reduces the total infrastructure requirement for the DoD.
- The digitization on non-electronically formatted data, and the transformation of the digital into common operating environment.
- The development of user-based search and retrieval tools based on the users defined requirements and the DKE environment.
- To improve on the DoD's ability to reduce acquisition/production lead-time related to the conversion of input data and in making information accessible to the user community.
- Reduce FOIA request cost to the customer and better serve the public.

Background

Headquarters U. S. Army Corps of Engineers has mandated the implementation of Corps of Engineers Electronic Document Management System (CEEDMS) by all USACE activities by September 2001. The CEEDMS project will consist of COTS Electronic Document Management Software that will incorporate document management, workflow, and recordkeeping. A CEEDMS Task Force has been formed to finalize the functional and technical specifications; develop a Cost-Benefit Analysis; obtain a Corps-wide requirements contract; define required CEEDMS standards; develop command management guidance; and develop a "typical" deployment plan for CEEDMS.

CEEDMS requirements contract will be in place by summer 2000. CEEDMS hardware, software, and services will be available for purchase at this time. Implementation of CEEDMS hardware and software will begin in August 2000. Complete implementation of CEEDMS is mandatory for all activities by the end of September 2001.

In conjunction with the implementation of the CEEDMS, existing hard-copy documents will need to be scanned, indexed, and entered into CEEDMS. It is estimated that 491 million documents will need to be scanned and indexed for insertion into CEEDMS, with approximately 4 percent of these documents being large (C-& E-sized) engineering drawings.

Level of Conversion to Support Mission

LETTER SIZED DOCUMENTS (8½ × 11 & 8½ × 14) – All hard-copy documents of this type will be evaluated to determine if the document needs to be retained. Once this procedure is complete, the documents that need to be kept will be scanned, indexed, and entered into CEEDMS. It is estimated that approximately 464.5 million documents of this type will need to be entered into CEEDMS. At a minimum, files will be converted to a pdf file format that

supports embedded ASCII strings. Minimum requirements for scanning the documents will include embedding the document title into the pdf file. This will enable the use of the document outside of CEEDMS with a text search engine. Vector conversions, Optical Character Recognition, and conversions to engineering models will be accomplished on an as needed basis. However, in addition to the desktop conversions, the need to perform this function may be determined prior to contract award, and provided in the contract specification by each field office. All documents will need keyword indexing performed for insertion into CEEDMS.

ENGINEERING DRAWINGS (C-& E-Sized) – All hard-copy documents of this type will be evaluated to determine if the document needs to be retained. All engineering drawings being retained will be scanned, indexed, and entered into CEEDMS. It is estimated that approximately 19.5 million documents of this type will need to be entered into CEEDMS. Scanned files should be saved in a suitable raster format such as TIFF or CALS that support embedded ASCII strings. Minimum requirements for scanning the documents will include embedding the document title into the TIFF or CALS file. Vector conversions, Optical Character Recognition, and conversions to engineering models will be accomplished on an as needed basis. However, in addition to the desktop conversions, the need to perform this function may be determined prior to contract award, and provided in the contract specification by each field office. All documents will need keyword indexing performed for insertion into CEEDMS.

Storage Plans

All documents will be processed by the ADCS contractor and entered into an ADCS CEEDMS storage area for Corps quality assurance. A Quality Assurance Team (QAT) will be formed at each Corps Activity. The QAT will ensure the integrity of the electronic documents before making them available. Once the electronic data is approved, it will be released by the QAT for Corps use within CEEDMS. CEEDMS will consist of an Oracle server, one or more repository servers, and a full text retrieval server at each field site. Although the CEEDMS is not currently in place, it is anticipated that the system will require at least one and probably several repository servers at each site. The various repository servers associated with the CEEDMS will be utilized to store the electronic documents. Storing the electronic documents in CEEDMS on repository servers will provide the ability for all Corps employees to access all Corps documents to which they have permissions.

Approach

The CEEDMS Implementation Schedule calls for the CEEDMS Task Force to test the software during the fourth quarter of FY-2000. The schedule also calls for the Corps-wide implementation of CEEDMS in the order of document management, recordkeeping, and workflow during the first, second, and third quarters of FY2001, respectively. Completion of implementation is mandated by Fourth quarter FY2001.

In order to take full advantage of the scanning initiatives, the CEEDMS Task Force will perform an ADCS test at CEEDMS Task Force support site (Huntington District) immediately following the testing of the CEEDMS software package. The pilot test will be initiated during the third quarter of FY2000 and will conclude at the end of FY2000. The implementation of the field portion of this ADCS project will begin at the end of implementation of the Document Management Phase, of CEEDMS, beginning Second Quarter FY-2001. ADCS for all field sites

will be initiated during the Second Quarter of FY2001 and will be completed by the end of fourth quarter of FY2005. Location of the scanning (at Corps field site or at ADCS contractor site) will be determined prior to contract award, and provided in the contract specification by each field office.

Deliverables

All deliverables will be provided as electronic documents, registered in CEEDMS via ADCS Contractor Client interface. The quantity various data types are shown below by field site. Estimated quantity of source hardcopy documents is shown below by field site. Actual quantities per data type along with data specification will be provided at contract time to the ADCS Contractor from each field office.

Table D-1 – Current Document Estimate Per Field Site

Organization	8½×11 & 8½×14 TIFF/PDF Pages	E & C Sized TIFF/CALS Pages
LRH (Huntington) Test	7,500	2,500
MDC	349,440	14,560
PODHQ	510,720	21,280
LRDHQ2	672,000	28,000
NWDHQ2	1,182,720	49,280
NADHQ	1,344,000	56,000
SADHQ	1,384,320	57,680
LRDHQ	1,585,920	66,080
SWDHQ	1,626,240	67,760
NWDHQ	1,680,000	70,000
SPDHQ	1,800,960	75,040
SAC	1,948,800	81,200
WRC	1,962,240	81,760
MVDHQ	2,136,960	89,040
CPW	2,150,400	89,600
UFC	2,701,440	112,560
HECSA	2,714,880	113,120
LRC	2,903,040	120,960
LRB	3,642,240	151,760
POF	3,830,400	159,600
SPN	3,884,160	161,840
POJ	4,072,320	169,680
SPA	4,287,360	178,640
CRREL	4,368,000	182,000
TAC	4,529,280	188,720
CERL	4,583,040	190,960
TEC	4,583,040	190,960
SWG	4,932,480	205,520
NAU	5,040,000	210,000
SAW	5,187,840	216,160
POH	5,443,200	226,800
POA	5,725,440	238,560
NAO	5,940,480	247,520

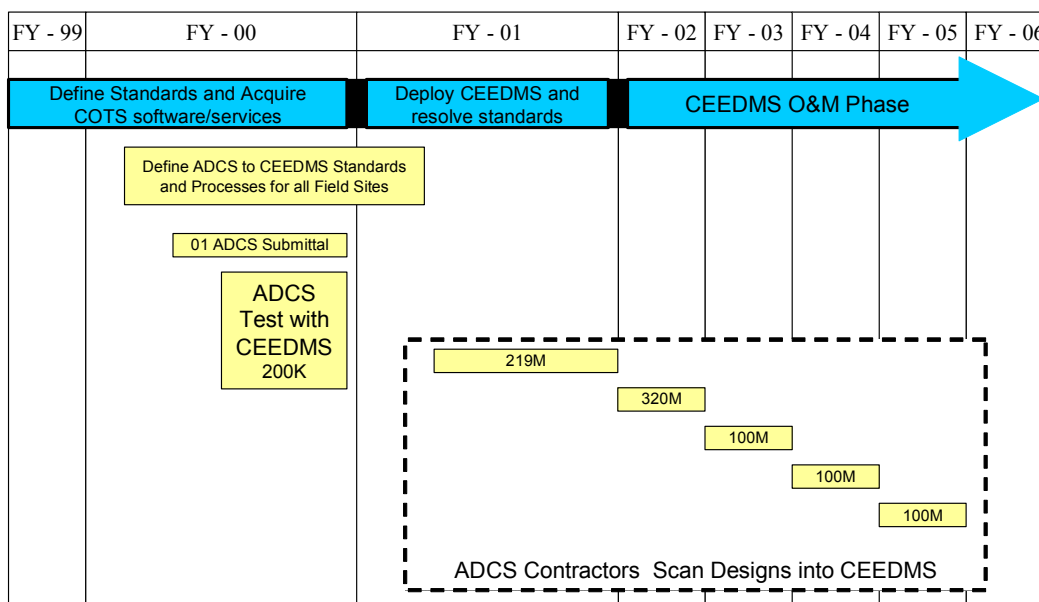
Table D-1 – Current Document Estimate Per Field Site

Organization	8½×11 & 8½×14 TIFF/PDF Pages	E & C Sized TIFF/CALS Pages
LRE	6,773,760	282,240
NAP	7,123,200	296,800
NAE	7,674,240	319,760
HNC	7,674,240	319,760
MVM	7,969,920	332,080
NWW	8,077,440	336,560
NAN	8,937,600	372,400
MVP	9,488,640	395,360
SWL	10,147,200	422,800
MVS	10,268,160	427,840
NWS	10,402,560	433,440
SAJ	10,631,040	442,960
SPL	10,671,360	444,640
LRN	10,967,040	456,960
HQUSACE	3,874,129	336,881
LRP	11,760,000	490,000
NWK	11,854,080	493,920
MVR	11,894,400	495,600
SWT	12,458,880	519,120
LRH	12,633,600	526,400
SWF	13,130,880	547,120
SPK	13,762,560	573,440
SAS	14,434,560	601,440
LRL	15,294,720	637,280
NWP	16,141,440	672,560
WES	16,558,080	689,920
MVK	16,625,280	692,720
NAB	16,692,480	695,520
MVN	16,786,560	699,440
NWO	17,377,920	724,080
SAM	17,740,800	739,200
TOTALS	464,537,629	19,533,381

SCOPE:

The project will require approximately 60 months to complete. This includes the time utilized in testing, documenting lessons learned, awarding the contracts, and doing bulk conversions for all field offices. Existing ADCS contractors will be used with Corps SOWs provided by each field office. The estimated schedule includes the following milestones:

Figure D-1 – 1999 EDMS and ADCS Project Schedule



Estimated Cost

The estimate to support this project was prepared using parametric estimating practices. The estimated cost of this project was based on several scanning needs analyses performed during the implementation plan phases of the Corps of Engineers Electronic Recordkeeping Information System (CEERIS). The quantity of documents to be scanned was gathered from several sites. The proposed number of documents to be scanned was then divided by the total number of full time employees (FTEs). From this we were able to predict the number of documents to be scanned for each District based on the number of FTEs working at each Corps organization. The average percentage of Drawings for the site was determined, and then used to determine the estimate of large-format scanning to be performed. The summary of the total costs, total costs by FY, and the costs by field sites are exhibited in the tables below.

Table D-2 – ADCS Project Total Cost Summary

U. S. Army Corps of Engineers EDMS Data Conversion Project = 850M		
Activity	Work Performed By	Estimated Cost
Bulk Conversion of Letter Sized (8 1/2 X 11 & 8 1/2 X 14)	Contractor - Competed	\$581M
Bulk Conversion of Engineering Drawings (C&E Sized)	Contractor - Competed	\$117M
Quality Assurance	QAT – Corps Field Sites	\$140M
Total Estimated Conversion Cost		\$838M

Table D-3 – ADCS Project Total Cost Per FY

	FY00	FY01	FY02	FY03	FY04	FY05
Proposed Project Funds	200K	218M	320M	100M	100M	100M
Minimum Required Project Funds	100K	130M	192M	60M	60M	60M

Table D-4 – ADCS Project Cost Per Site

Organization	8 1/2 X 11 & 8 1/2 X14 TIFF/PDF Cost	E & C Size TIFF Cost	Total Cost per Site
LRH (Huntington) Test	\$9,375.00	\$15,000.00	\$24,375.00
MDC	\$436,800.00	\$87,360.00	\$524,160.00
PODHQ	\$638,400.00	\$127,680.00	\$766,080.00
LRDHQ2	\$840,000.00	\$168,000.00	\$1,008,000.00
NWDHQ2	\$1,478,400.00	\$295,680.00	\$1,774,080.00
NADHQ	\$1,680,000.00	\$336,000.00	\$2,016,000.00
SADHQ	\$1,730,400.00	\$346,080.00	\$2,076,480.00
LRDHQ	\$1,982,400.00	\$396,480.00	\$2,378,880.00
SWDHQ	\$2,032,800.00	\$406,560.00	\$2,439,360.00
NWDHQ	\$2,100,000.00	\$420,000.00	\$2,520,000.00
SPDHQ	\$2,251,200.00	\$450,240.00	\$2,701,440.00
SAC	\$2,436,000.00	\$487,200.00	\$2,923,200.00
WRC	\$2,452,800.00	\$490,560.00	\$2,943,360.00
MVDHQ	\$2,671,200.00	\$534,240.00	\$3,205,440.00
CPW	\$2,688,000.00	\$537,600.00	\$3,225,600.00
UFC	\$3,376,800.00	\$675,360.00	\$4,052,160.00
HECSA	\$3,393,600.00	\$678,720.00	\$4,072,320.00
LRC	\$3,628,800.00	\$725,760.00	\$4,354,560.00
LRB	\$4,552,800.00	\$910,560.00	\$5,463,360.00
POF	\$4,788,000.00	\$957,600.00	\$5,745,600.00
SPN	\$4,855,200.00	\$971,040.00	\$5,826,240.00
POJ	\$5,090,400.00	\$1,018,080.00	\$6,108,480.00
SPA	\$5,359,200.00	\$1,071,840.00	\$6,431,040.00
CRREL	\$5,460,000.00	\$1,092,000.00	\$6,552,000.00
TAC	\$5,661,600.00	\$1,132,320.00	\$6,793,920.00
CERL	\$5,728,800.00	\$1,145,760.00	\$6,874,560.00
TEC	\$5,728,800.00	\$1,145,760.00	\$6,874,560.00
SWG	\$6,165,600.00	\$1,233,120.00	\$7,398,720.00
NAU	\$6,300,000.00	\$1,260,000.00	\$7,560,000.00
SAW	\$6,484,800.00	\$1,296,960.00	\$7,781,760.00
POH	\$6,804,000.00	\$1,360,800.00	\$8,164,800.00
POA	\$7,156,800.00	\$1,431,360.00	\$8,588,160.00
NAO	\$7,425,600.00	\$1,485,120.00	\$8,910,720.00
LRE	\$8,467,200.00	\$1,693,440.00	\$10,160,640.00
NAP	\$8,904,000.00	\$1,780,800.00	\$10,684,800.00
NAE	\$9,592,800.00	\$1,918,560.00	\$11,511,360.00
HNC	\$9,592,800.00	\$1,918,560.00	\$11,511,360.00
MVM	\$9,962,400.00	\$1,992,480.00	\$11,954,880.00
NWW	\$10,096,800.00	\$2,019,360.00	\$12,116,160.00
NAN	\$11,172,000.00	\$2,234,400.00	\$13,406,400.00

Table D-4 – ADCS Project Cost Per Site

Organization	8 1/2 X 11 & 8 1/2 X14 TIFF/PDF Cost	E & C Size TIFF Cost	Total Cost per Site
MVP	\$11,860,800.00	\$2,372,160.00	\$14,232,960.00
SWL	\$12,684,000.00	\$2,536,800.00	\$15,220,800.00
MVS	\$12,835,200.00	\$2,567,040.00	\$15,402,240.00
NWS	\$13,003,200.00	\$2,600,640.00	\$15,603,840.00
SAJ	\$13,288,800.00	\$2,657,760.00	\$15,946,560.00
SPL	\$13,339,200.00	\$2,667,840.00	\$16,007,040.00
LRN	\$13,708,800.00	\$2,741,760.00	\$16,450,560.00
HQUSACE	\$4,842,661.50	\$2,021,284.80	\$6,863,946.30
LRP	\$14,700,000.00	\$2,940,000.00	\$17,640,000.00
NWK	\$14,817,600.00	\$2,963,520.00	\$17,781,120.00
MVR	\$14,868,000.00	\$2,973,600.00	\$17,841,600.00
SWT	\$15,573,600.00	\$3,114,720.00	\$18,688,320.00
LRH	\$15,792,000.00	\$3,158,400.00	\$18,950,400.00
SWF	\$16,413,600.00	\$3,282,720.00	\$19,696,320.00
SPK	\$17,203,200.00	\$3,440,640.00	\$20,643,840.00
SAS	\$18,043,200.00	\$3,608,640.00	\$21,651,840.00
LRL	\$19,118,400.00	\$3,823,680.00	\$22,942,080.00
NWP	\$20,176,800.00	\$4,035,360.00	\$24,212,160.00
WES	\$20,697,600.00	\$4,139,520.00	\$24,837,120.00
MVK	\$20,781,600.00	\$4,156,320.00	\$24,937,920.00
NAB	\$20,865,600.00	\$4,173,120.00	\$25,038,720.00
MVN	\$20,983,200.00	\$4,196,640.00	\$25,179,840.00
NWO	\$21,722,400.00	\$4,344,480.00	\$26,066,880.00
SAM	\$22,176,000.00	\$4,435,200.00	\$26,611,200.00
TOTALS	\$580,672,036.50	\$117,200,284.80	\$697,872,321.30

SUPPORT DOCUMENTS

- CEEDMS Schedule
- Corps FTE Summary Spreadsheet
- Detailed ADCS Cost Estimate
- CEEDMS Task Force Roster

Appendix E. Sample Conversion Worksheet

PROJECT		CONTRACT NUMBER		REQUIREMENTS WORKSHEET	
ORGANIZATION		DATE			
The following questions are designed to help Program Managers compile the information required to answer the questions relating to document conversion.					
No.	Question	User Input			
1	Functional Type of Documents?	Mechanical			
		Electrical			
		Architectural/Civil			
		GIS Mapping			
2	Number of Engineering Documents?				A
					B
					C
					D
					E
					F
					G
					H
					J
					K
3	How Will Converted Document Be Used?	Design			
		Design Analysis			
		Production			
		Procurement			
		Support/Maintenance			
		Technical Manuals			
		Illustrations			
		Specifications			
4	Type of Access of Documents?	View Only			
		Comment/Annotate			
		Process/extract/transform			
		Edit/update/maintain			
		Archive			
5	Identify User Infrastructure (Operating System)	DOS			
		Window NT/9X			
		UNIX			
6	Identify Existing Input Format of All Documents	Paper			
		Mylar			
		Aperture Cards			
		Microfiche			
		TIFF			
		CALS Type 1			
		JEDMICS			
7	Delivery Media to Contractor (If Electronic)	CD-ROM			
		Network (Govt.)			
		Disk (ZIP Drive)/			
		Tape (DAT Drive)			
		Other			

PROJECT		CONTRACT NUMBER		REQUIREMENTS WORKSHEET	
ORGANIZATION		DATE			
The following questions are designed to help Program Managers compile the information required to answer the questions relating to document conversion.					
No.	Question	User Input			
8	Delivery Method to Contractor	US Mail - First Class Ground			
		Commercial Carrier			
		Hand Delivery			
		Bonded Messenger			
		Electronic Transfer			
9	Converted Output Document Type	Mechanical 2D CAD Capable			
		Mechanical 2D CAD Perfect			
		Mechanical 3D CAD Perfect			
		Electrical 2D CAD Capable			
		Electrical 2D CAD Perfect			
		Electrical Schematic			
		Architectural 2D CAD Capable			
		Architectural 2D CAD Perfect			
		Architectural 3D CAD Perfect			
		Civil 2D CAD Capable			
		Civil 2D CAD Perfect			
		Civil 3D CAD Perfect			
		GIS/Mapping 2D CAD Capable			
GIS/Mapping 2D CAD Perfect					
GIS/Mapping 3D CAD Perfect					
10	Conversion Output Data Format	Vector Formats			
		C4		DWG	
		Native		IGES	
		DXF		Step AP 203	
		IGES		DGN	
11	Metadata	JEDMICS/CDEX			
		DoD			
		Product Data (PM)			
12	Output Media to Program Manager	CD-ROM			
		Disk (ZIP Drive)			
		Tape (DAT Drive)			
		Network (ADCS WEB Site)			
13	Method of Labeling Output Media	Printed Label			
		Print/ID			
		Electronic Tag			
		URL Tag			
		Other			
14	Delivery Method to Program Manager	US Mail - First Class Ground			
		Commercial Carrier			
		Government Personnel			
		Bonded Messenger			
		Electronic Transfer			
15	Disposition of Originals	Return Document			
		Shred Document			
		Other			

Appendix F. Standards and Estimates

F.1 Standards

The average computer file sizes for many types of scanned (digitized) documents are listed and explained below.

F.1.1 Introduction

Round Numbers

All of the estimates are chosen to be representative of the scanned document images and to produce easily useable, round figures when multiplied. For example, 50,000 bytes (50 kilobytes) is close to the average size of a scanned page and also yields an estimate of exactly one million bytes (a megabyte) for 20 pages, exactly one billion bytes (a gigabyte) for 20 thousand pages, and exactly one trillion bytes (a terabyte) for 20 million pages.

Over-Estimation

These estimates have been chosen to tend toward over-estimation rather than under-estimation of storage requirements. The estimates therefore provide a small safety margin. It is best to make certain that all estimates and estimating procedures tend toward conservative estimates. Then, when all assumptions are factored in together, aggregated, the overall estimate is conservative (safe).

Metric Round Numbers

The United States became a metric country in 1866 (Act of July 28, 1866; 14 Stat. 339) when the US customary measures were defined in metric terms. For example, an inch is defined to be exactly 25.4 millimeters. While an inch, as a measure, is thus hard metric, it is not a round metric number. When physical items are hard metric, the quantity and size of the items are in round metric numbers. In this paper, US customary measures are converted to metric sizes and quantities by applying the above methods for constructing round numbers and ensuring a slight over estimation. These methods are applied to the quantities and sizes of documents listed once for the US customary units, and a second, independent time for metric units. The resulting metric measures are then chosen to be appropriate to the corresponding US customary measures. This means that the conversion process produces similar conceptual values, but almost never produces numerically or physically equal values (*e.g.*, -40 degrees Celsius = -40 degrees Fahrenheit) for the quantity or size of a given item.

Variances from Average

Making the assumption that your documents are similar to the industry average documents usually produces very small variances. Because the cost of storage is very low as a percentage of overall system cost, and is dropping rapidly, an error of a few percent in an estimate has very little effect on the overall system cost. If round estimates speed up the understanding and discussion process, the benefit of rounding far out-weighs the cost of the slight variances.

After one percent of the documents have been scanned into a system, an actual average page image size can be calculated. This actual average page image size will provide the small correction necessary to adjust previous estimates. This is the system sizing method used in almost all system implementations.

The Value of Estimating

Managers need ball park figures to size a document management system, to identify what system elements are bigger (or smaller) than a breadbox (system or system component). Managers need a rough order of magnitude (ROM) estimate. An order of magnitude is a power of 10 so that a rough order of magnitude estimate (ROM) is one in which the largest reasonable estimate is about ten times the size of the smallest reasonable estimate.

The accompanying list of document sizes is intended to assist in creating ROM estimates of storage requirements. Possible system decision outcomes of making these estimates are: “There will be no problem (the system is a little too large),” “We will never be able to afford the system,” and “We budgeted the right amount.”

If everyone uses the same estimates, it is easier to discuss and compare document imaging systems. Managers can also benefit from reports and articles describing previously implemented systems. Because the estimates are industry standard, less time can be spent evaluating estimating methods, and more time can be spent understanding how the system will be used and whether the system design will accommodate the planned use.

The Difference between Pages and Documents

The document imaging industry has blurred, and nearly eliminated, the difference in meaning between the words “page” and “document.” In making systems seem as large as possible, every stored page is called a “document” in marketing materials. To permit meaningful discussion of pages (individual sheets of paper) versus documents (one or more pages that belong together), always separately list the number of documents, the number of pages, and the average number of pages per document.

The Difference between Pages and Page Images

These estimates are counted by page images. To avoid confusing the number of pages (individual sheets of paper) with the number of page images (sides of a page containing information), always separately list the number of pages and the number of page images.

F.1.2 Final Adjustment Factor

If your storage facilities are tightly packed or are otherwise different than these estimates, you can apply a correction factor to adjust for you facility’s differences. For example, a tightly packed facility might have an adjustment factor of 1.1 because there might be ten percent more pages per linear foot or drawer than in the industry standard density.

The adjustment factor is not required for rough, order-of-magnitude (ROM) estimates. Also, the adjustment factor is easy to apply at any stage in the process. If after weeks of work, a storage estimate of 100 gigabytes is arrived at, the decision to use an adjustment factor can be made, and

the storage estimate can simply be adjusted to 1.1 times 100 gigabytes, yielding an estimate of 110 gigabytes.

F.1.3 Units of Measure (Digital)

One byte (B) is defined as the set of bits used to represent 1 character. Commonly, 1 byte = 8 bits (b). (When speaking of both units in the same document, “byte” and “bit” are often best spelled out.) An eight-bit ASCII byte can therefore represent any of 256 (2^8) different characters. By comparison, one Unicode byte = 16 bits = 1 character. A 16-bit Unicode byte can represent any of 65,536 different characters, enough capacity to include most of the world’s written characters in the same character set.

Computer storage is almost exclusively given in bytes and transmission speeds are almost exclusively given in bits. Diligent care in using the two terms will result in plans that are neither eight times too large nor eight times too small.

- 1,000 bytes = 1 kilobyte (KB) (exactly 1 thousand in common and legal usage; in computer terms, $1 \text{ KB} = 2^{10} = 1,024$ bytes); 1,000 KB = 1 megabyte (exactly 1 million in common and legal usage; in computer terms, $1 \text{ MB} = 2^{20} = 1,048,576$ bytes); 1,000 MB = 1 gigabyte (billion, GB); 1,000 GB = 1 terabyte (trillion, TB); 1,000 TB = 1 petabyte (quadrillion, PB); 1,000 PB = 1 exabyte (quintillion, EB); 1,000 EB = 1 zettabyte (sextillion, ZB); 1,000 ZB = 1 yottabyte (septillion, YB).
- 1 millisecond (ms) = .001 second; 1 microsecond (μs) = .001 ms, 1 nanosecond (ns) = .001 μs ; 1 picosecond (ps) = .001 ns; 1 femtosecond (fs) = .001 ps; 1 attosecond (as) = .001 fs; 1 zeptosecond (zs) = .001 as; 1 yoktosecond (ys) = .001 zs.
- 1 Hertz = 1 cycle per second (used in computer terms to denote 1 clock cycle in a computer which corresponds roughly to 1 instruction execution.). One thousand cycles per second is 1 kiloHertz (KHz); each cycle of such a signal or action is a millisecond long. 1,000 KHz = 1 megaHertz (MHz), etc. Because light travels about 300 megameters (Mm) in 1 second and has a wavelength of about 400 nm for blue light (about 700 nm for red light), the frequency of light is about 750 THz for blue light (about 430 THz for red light). This is because the speed of any electromagnetic wave (*i.e.*, the speed of light, a constant) = wavelength \times frequency.

Why is a computer MB not exactly one million bytes? Because computers use binary arithmetic and the closest round number computers have that is near one million is two to the twentieth power. Two to the twentieth power is equal to 1,048,576. This is why many computer program displays show an exact number of bytes beside a seemingly smaller number of megabytes. (For example, see any Windows disk or file properties display.) Similarly, a computer kilobyte is not exactly one thousand bytes, but is two to the tenth power or 1,024 bytes.

Because of lawsuits over the meaning of kilobyte, megabyte, etc., only the metric meanings (based on units of 1 thousand) are permitted in commercial discussions of storage capacities. The computer based terminology continues in use when discussing computer configurations because computers actually use equipment based on binary counting methods. The computer units must always be converted into commercial metric (1 thousand based) units before commercial

discussions take place. Document imaging and document management discussions fall into the category of commercial discussions.

F.1.4 Transmission Time Standards

There are several common communications line types available. The speed of each line type in bits per second and page images per second is given along with a rough estimate of the monthly cost of a local connection (two or three miles). This will help in confirming that the speed of access desired is possible over the communications lines proposed.

Modem = 56 kilobits per second (kbps) = 3 pages per minute (about US\$30.00 per month)
ISDN (2 voice channels) = 128 kbps = 10 pages per minute (about US\$100.00 per month)
Cable (TV) Modem = 500 kbps = 1 page per second (about ~ US\$50.00 per month)
T1 (24 voice channels) = 1.544 megabit per second (Mbps) = 3 pages per second (about US\$1,000 per month)
Ethernet (CSMA/CD) = 1 Mbps (effective) or 10 Mbps (nominal) = 2 pages per second
OC3 ATM (Optical Carrier, Asynchronous Transfer Mode) = 155 Mbps = 300 pages per second
OC192 (SONET: Synchronous Optical NETwork fiber) = 10 gigabits per second (Gbps) = 20,000 pages (2 file cabinets) per second
Dense Wavelength Division Multiplexing (DWDM) with OC192 = 320 Gbps = 64 file cabinets per second
Optical Carrier Frequency (1,300 nm) = 230 THz (about 20,000 cycles are used for every OC192 bit transmitted)

F.1.5 Pages, File Cabinets, Boxes, and Linear Feet

Using an estimate of 2,500 pages per file drawer and four file drawers per file cabinet, one can estimate that scanning one four drawer file cabinet full of documents (ten thousand single sided pages) will fill one CD ROM disc. Similarly, the scanned contents of two file cabinets will fill one gigabyte of magnetic disk storage.

With these figures, a simple count of the file cabinets in an organization will produce an estimate of the amount of storage required. At an even coarser level, if two file cabinets are assumed for each employee, the number of gigabytes required is equal to the number of employees.

Files in file cabinets and on linear feet of open shelving are assumed to have some open space for file growth and for ease of access. The number of pages estimated for these two storage methods takes this into account.

Standard Record Storage Cartons and Fan-Folded Computer Output

A standard records storage carton (box) is about 12 inches wide by 15 inches long by 9½ inches deep. It is designed to store letter size documents in manila folders against the 12-inch side and legal documents in legal folders against the 15-inch side. These boxes are assumed to be tightly packed because documents in boxes are placed there for storage. Access to documents in boxes is assumed to be less frequent than to documents stored in active file cabinets and on open shelves. Therefore a standard records storage carton can be assumed to hold 2,500 pages, the same as one file drawer. For double-length boxes, those measuring 12×30 inches instead of the more standard 12×15 inches, multiply the number of pages per box by 2 and assume 5,000 pages per box instead.

Standard fan-folded, greenbar, tractor fed, 11 × 14 inch computer paper can be placed flat on the bottom of the standard records storage carton. Because the mainframe style programs that produce fan-folded greenbar output use very simple typography, these pages tend to be simple and compress well, to about the same size as an average letter size page. Fan-folded documents are on fine paper and must be handled carefully or they quickly become unmanageable. For these reasons, the nine or so inches that will fit in a standard records storage carton are fairly dense with information and are about equivalent to the same 2,500 sheets of letter size paper.

F.1.6 Digital Media Capacity

Table F-1 – Storage Capacities of Standard Storage Media

Storage Media	Capacity (GB)	Transfer (MB/sec)	Seek (ms)	Cost (\$/MB)
Magnetic Hard Disk	9 and up	7-20	~11	.15-.25
Removable Magnetic	1	~5	~11	.12
Magneto Optical	4.6	~4	~19	.05-.11
WORM	1.3-4.6	~4	~19	.05-.11
CD-R (r/o)	.650	0.3-0.8	~200	~.01
CD-ROM (r/o)	.650	0.3-0.8	~200	~.002
DVD (r/o)	4.7-17	0.4-1.0	~200	<.001

F.1.6.1 CD-ROM

The CD (Compact Disc) development was funded by the music industry. The peak capacity of a CD-ROM (a CD holding computer data) is about 650 MB, but CD-ROMs are rarely filled to peak capacity. A convenient estimate is to assume each CD holds 500 MB in actual practice. This is consistent with the goal of erring on the conservative side and of slightly overestimating the storage media required at every step. With these estimates, one CD can therefore be assumed to hold the scanned contents of one standard file cabinet, which produces a simple, round number, estimate.

1 CD-ROM (Read Only Memory) = 500 – 650 MB
--

F.1.6.2 DVD-ROM

Development of the DVD (usually Digital Video Disc) has been funded by the video industry, the movie industry, the music industry, and the computer industry. The DVD is expected to unify the PC, TV, telephony, and document management. The DVD has several versions and options. The table below shows the capacity of each version.

The estimate of 10 file cabinets per DVD-ROM (DVD holding computer data) is very conservative. It assumes considerable overhead for storing indexes, and gives a large amount of weight to the goal of creating round number estimates, using a figure of 10 rather than 12 file cabinets per DVD.

A DVD supports 5 channel (theater quality) surround sound, 96 KHz / 24 bit audio, 8 language tracks, 32 subtitle tracks, and about 135 minutes of high quality (720 horizontal lines) video (long enough to accommodate 94% of all movies) on each of 4 layers. DVDs support runtime editing so that all ratings of a movie are on the same DVD; scenes can be skipped as the DVD is

played. The file format is ISO 13346 UDF (Universal Disc Format) which harmonizes all CD recording standards including ISO 9660. A future technology, third-generation blue lasers, may yield a 40 gigabyte DVD-ROM sufficient to support single-disc HDTV-quality recordings.

1 DVD-WORM: (Write Once, Read Many) (2 sided, 1 layer per side) = 7.9 GB (3.95 GB per side).
1 DVD-RW: (Re-Writable) (2 sided, 1 layer per side) = 5.2 GB
1 DVD-ROM (Read Only Memory) (2 sided, 2 layers per side) = 17 GB

DVD Audio

DVDs can be used to record audio only, with no video. A DVD audio disc may nevertheless include still images. DVD audio is different than the audio used on a DVD video disc.

The DVD audio standard is for up to 6 channels, a sampling rate of 48, 96, or 192 KHz, and a sample size of 16, 20, or 24 bits. With 24 bit samples taken at a 192 KHz rate, this provides a 96 KHz frequency response and a 144 dB dynamic range. DVD audio can also provide for a lossless audio compression of about 2 to 1 which would have a playing time of 120 to 140 minutes for two- channel 192 KHz / 24 bit recordings for a single layer. Each DVD disc can have up to 4 layers, 2 layers per side.

DVD audio includes various still image modes for synchronized lyrics, navigation, etc. DVD audio allows up to 16 still graphics per track and a set of limited transitions.

The audio used in DVD video can also be used without the video. This produces a stereo, DVD quality, play time of over 55 hours at 192 Kilobits per second (compressed) for a single layer and over 200 hours for a 4 layer DVD disc. Lower quality sound can be recorded as computer files on a DVD for much longer play times. At a compressed audio rate of 16 Kilobits per second (in the low range of telephony quality), this is 9 million seconds, 150 thousand minutes, 2,500 hours, 100 days, 15 weeks, or 3 months of audio on a 4 layer DVD disc.

F.1.7 Pixel Sizes and Pixels per Image

Most document imaging resolution measures are in pixels (PICTure ELEment) per inch (or per mm - millimeter), and are commonly referred to as dpi (Dots per Inch) or dpmm (Dots per mm). Most motion-picture and still-photographic resolution measures are in pixels per image. This is most commonly seen in the 525 lines of NTSC (National Television System Committee), 625 lines for PAL (Phase Alternating Line) and SECAM (Sequential Couleur Avec Memoire or Sequential Colour with Memory), resolution of television images. No matter how physical large or small an NTSC television image is displayed, there are only 525 lines of vertical resolution (480 viewable). The computer equivalent of this is 640 by 480 pixels in a standard computer image.

In pixels per image the horizontal resolution is given first. If the horizontal dimension is larger than the vertical dimension in pixels, the image is said to be landscape, if the horizontal is smaller, the image is said to be portrait.

Pixel size is based on the image or object scanned, for example $\frac{1}{300}$ inch (square) pixels scanned from a letter size page scanned at 300 dpi. If the image has been microfilmed at a reduction of 12 \times then the pixels scanned on the microfilm are scanned at 3600 dpi ($\frac{1}{3600}$ inch square) to have an effective resolution of 300 dpi relative to the original letter size page. If the pixels are

displayed on a monitor that has a resolution of 100 dpi, then the pixels are 100 dpi ($1/100$ inch square), and have been enlarged, along with the document, by a factor of 300 percent. If the pixels are combined, 9 pixels to 1 pixel (3 pixels to 1 pixel in both dimensions of the two dimensional image), to display the letter size document at a 1 to 1, normal size, on a 100 dpi monitor, then the pixel size is increased and the image resolution is decreased.

Resolution

For snapshot photographs and typed documents, image files created by scanning at 200, 300, and 400 dpi (dots per inch) all have approximately the same information content as the original image. The higher resolutions merely increase the redundancy in the image file. It is this redundancy that compression removes. In general, higher resolution scans of an image are slightly larger than lower resolution scan of the same image because higher resolution scans pick up more noise (pseudo-information). (Noise is something like digital dirt on the image.)

This variation between the compressed image sizes of different resolutions is within the variation range of document image sizes in general. In almost all cases, measuring the actual sizes of the first one percent of scanned images will easily adjust for this variation without requiring significant system changes.

Figure F-1 – Creating an Image File from the Printed Page

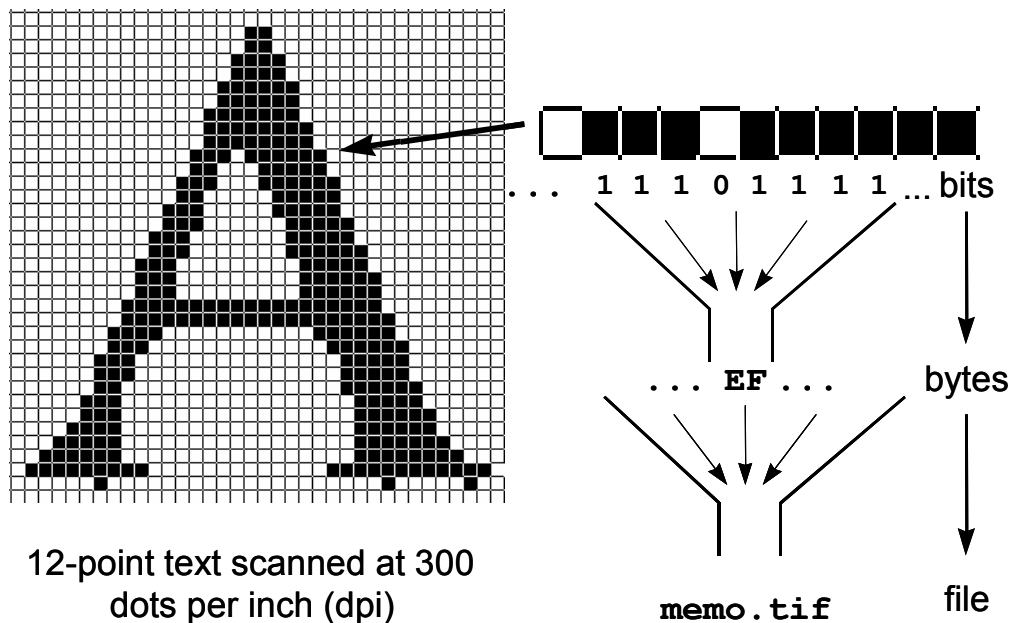


Figure F-2 – Resolution and File Size Comparison

Scan Resolution	Image Sample	TIFF Group 4 File Size
100 dpi	Digital	29 KB
200 dpi	Digital	55 KB
300 dpi	Digital	83 KB
400 dpi	Digital	110 KB
600 dpi	Digital	169 KB

When speaking of pixels per image, however, the number of pixels, and the amount of information in the pixels, stays fixed no matter what size the image is when reproduced. This relationship can be clearly seen in a store selling many sizes of television sets. If the same demonstration video clip is being shown on every television set, then every set has the same picture, with all the same picture information, no matter what size the television set is.

Pixels do not have a size in the computer or when they are stored digitally. A raster of digitally stored pixels is just an array of numbers. This array of number carries all the image information.

Microprocessor and RAM (Random Access Memory) Design Rules (Pixel Size)

Semiconductors are made using digital photographic techniques (pixels). Recently, microprocessor production processes were improved from .25 micron (.25 μm , micrometer) (250 nm, nanometer) design rules to .18 μm (180 nm) design rules. This means that the pixel size for semiconductor devices is now slightly less than $\frac{1}{5}$ micron (200 nm). A micron is $\frac{1}{10^6}$ meters. Using 200 nanometer (nm) pixels and assuming $\frac{1}{25}$ of the area was used for active transistors, a 1 millimeter (mm) square area (about the size of the head of a pin) could hold 25 megaPels (25 million pixels) and 1 million transistors.

The smallest practical pixel would be a pixel used as part of a halftone dot that represented the edge of the path of a sub-atomic particle, such as a neutrino. To create a smooth path in a specific color, a printed resolution of 2540 dpi (100 dpmm) would be used. Assuming a 1 ym (yoktometer) wide path, rendered as a 10 mm wide path, the width represented by each pixel would be .001 ym. For a superstring (2×10^{-35} m wide), the pixel width would be $2 \div 10^{12}$ ym.

Aerial Photography, Digital Orthophotography, and Remote Sensing

Aerial photography uses photographs taken from the air, recording the visible electromagnetic spectrum (light), as maps of geographic areas. Remote sensing includes photographs taken from the air and from beyond the atmosphere of areas on the earth and other celestial bodies, using

many segments of the electromagnetic spectrum including visible light, ultraviolet, infrared, and radar illumination. Digital orthophotography digitally rectifies the pixels of digitized aerial photographs into a continuous map, usually registered to a layer of a GIS (Geographic Information System).

For cities, 2 inch to 6 inch pixels are popular for digital orthophotography. A digital orthophotograph of a 500 square mile city using 6-inch pixels would have 4 pixels per square foot, 100 million pixels per square mile (there are approximately 25 million square feet per square mile), for a total of 50 gigaPels (50 billion pixels). Using 24-bit color and estimating a lossless three-to-one compression, this digital orthophographic image would require 50 gigabytes to store. If 2-inch pixels were used, a 500 square mile city would have 9 times as many pixels or 450 gigaPels, requiring 450 gigabytes to store using the same assumptions. Using 2-inch pixels a 50 square mile city would have 45 gigaPels, requiring 45 gigabytes to store using the same compression assumptions. The metric equivalents are 50 millimeter (mm) and 100 mm pixels which are respectively 400 and 100 to the square meter. For a 1 thousand square kilometer city this would be 100 gigaPels using 100 mm pixels and would require 100 gigabytes to store. Using 50 mm pixels for a 1 thousand square kilometer city, this would require 400 gigaPels requiring 400 gigabytes to store. A 100 square kilometer city, using 50 mm pixels would be imaged in 40 gigaPels which would require 40 gigabytes to store.

In digital orthophotography, in addition to color, each pixel has an associated z-axis value, the height of the pixel above sea level. When added to the x and y Cartesian coordinates of the pixel, the z values construct a digital terrain model over which the image can be mapped as a surface. This is similar to the way that images are created in virtual reality. By adding a t value, a 4 fourth dimension that represents a specific point in time, animations can be done telling a geologic story or the development history of a city.

In remote sensing (satellite imagery such as weather photographs or images for crop quality assessment or storm and flooding damage), an image of an area 1,000 kilometers square, using 100-meter pixels (pixels that are 100 meters by 100 meters), would contain 100 million pixels. Estimating a lossless three-to-one compression, this image would require 100 MB to store. The pixels used can be of any size. In astronomy, a single pixel can include an entire earth type planet (10,000 kilometer or 10 megameter pixels), a sun type star (1 million kilometers or 1 gigameter), or a galaxy (100,000 light-year pixels or approximately 1 zettameter pixels). The largest practical pixel is 400 yottameters square, where 400 yottameters is the estimated diameter of the observable universe.

F.1.7.1 Pixels per Image: Video Image Resolutions

Computer screen resolutions are chosen to have an aspect ratio (the ratio of width to height) of 4 to 3 (the “golden ratio” of the art world) and to have the number of pixels be an integer multiple of a power of 2.

Common display resolutions:

CGA (Color Graphics Adapter) 320 × 200 EGA (Enhanced Graphics Adapter) 640 × 350 VGA (Video Graphics Array) 640 × 480 Super VGA 800 × 600 XGA (Extended Graphics Array) 1024 × 768 XVGA (Extended VGA) and UVGA (Ultra VGA) 1024 × 768 Note SVGA, XVGA, and UVGA can also mean merely any display with greater resolution than 640 × 480.
1600 × 1200 high resolution document imaging workstation 1800 × 1440 high resolution document imaging workstation 2048 × 1536 high resolution grayscale document imaging workstation
Sun Microsystems 1152 × 900; Apple Macintosh 1152 × 870 (1152 = 2 ⁴ × 72). Some Sun Microsystems and Apple / Mac screen resolutions were chosen so that the actual screen resolutions were 72 dpi, to match the 72 points per inch used in typesetting.
DVD NTSC resolution is 720 × 480 and the DVD PAL/SECAM resolution is 720 × 576.
HDTV (High Definition TV) resolution is 1920 × 1200 (Sun Microsystems) at the 16:9 aspect ratio. The 1920 × 1200 resolution is designed to accommodate the NTSC-derived HDTV video resolutions of 1920 × 1080 and 1920 × 1035 and the PAL- and SECAM-derived HDTV video resolution of 1920 × 1152.

F.1.7.2 Pixels per Image: Single Frame (Still) Image Resolutions

Kodak PhotoCD resolutions are based on a 2:3 aspect ratio and an integer power of 2. The multiple of the base gives the number of pixels per image relative to the base image size in pixels. A Kodak PhotoCD contains five resolutions of each image: 1/16 Base through 16 Base. (The average compressed file size containing all five resolutions is about 5 MB per image.) A Kodak Pro PhotoCD contains the five resolutions for each image found on a PhotoCD plus a sixth 64 Base resolution.

1/16 base = 128 px × 192 px	thumbnail, index print on CD cover
1/4 base = 256 px × 384 px	largest Kodak size that is smaller than 480 × 640 for display on TV
1 base = 512 px × 768 px	
4 base = 1024 px × 1536 px	largest Kodak size that is smaller than 1920 × 1152 for HDTV
16 base = 2048 px × 3072 px	captures all the resolution on most 35 mm film images
64 base = 4096 px × 6144 px	for most film formats larger than 35 mm

F.2 Estimating Storage Requirements for Digitized Document Types

All image sizes are 1 bit per pixel raster images, black & white, and compressed, unless otherwise noted. In general, when compressed, the digital files for document images scanned in office quality color are about twice the size of document images scanned in a bi-tonal, black and white format, and then G4 compressed. In office quality color scanning, the scanned color differences aid in reading a document and in increasing the quality of Optical Character Recognition (OCR). Office quality color scanning is generally at a lower resolution than black and white scanning. Office quality color includes (subsumes) the process of grayscale scanning which can increase OCR accuracy when using low resolution scanning.

F.2.1 Scanned Letter Size Pages

1-Bit (Black and White)
1 scanned page (US letter or A4) (black & white, CCITT G4 compressed) = 50 KB 1 file cabinet = 4 file boxes/drawers = 10,000 pages = 500 MB = 1 CD-ROM 2 file cabinets = 1 GB 10 file cabinets = 1 DVD-ROM 2,000 file cabinets = 1 TB = 200 DVD-ROMs 1 file box/drawer (12" × 15" × 9.5"; 300 × 375 × 250 mm) = 2,500 pages = 2 linear feet = 125 MB 8 file boxes/drawers = 2 file cabinets = 20,000 pages = 16 linear feet = 1 GB 8,000 file boxes/drawers = 2,000 file cabinets = 20 million pages = 16,000 linear feet = 1 TB
Office Color
1 scanned page (US letter or A4) = 100 KB (office color, including grayscale, compressed) 1 file cabinet = 1 GB = 2 CD-ROMs 5 file cabinets = 1 DVD-ROM 1,000 file cabinets = 1,000 gigabytes = 1 TB; 1,000 file cabinets = 200 DVD-ROMs 1 file box (12" × 15" long × 9.5 deep) (2,500 pages) = 1 file drawer = 2 linear feet of files = 250 MB 4 file boxes = 8 linear feet = 1 file cabinet = 1 GB 4,000 file boxes = 8,000 linear feet = 1 TB
OCR/Text
1 text (non-raster; OCR or word-processor) page (US letter or A4) (word processor file format) = 5 KB

Legal Size

Legal size page images are not much larger than letter size page images when scanned and compressed. The pathological case is the 8½" × 25" contract set in 6-point type, created to be true to the statement that the entire contract is on just one page (which itself is almost always two sided). Even in these cases, a good system design can be arrived at using letter size page image estimates. Measuring the size of actual scanned pages after the first one percent of the legal size documents have been scanned will produce the same high level of accuracy produced by measuring the size of the first one percent of any type of scanned document images. With this foundation of industry de facto standards, adjustments can be made for even the worst pathological cases, and any desired degree of estimating accuracy can be achieved. By counting a standard record container that contains about 2,000 legal pages as having 2,500 sheets of letter size documents, the effect of the slight difference between legal and letter pages can be further reduced.

Index Storage Requirements and OCR Pages

Storage requirement estimates can be done more quickly by ignoring the storage requirements for OCR images and indices. This shortcut will not greatly affect the accuracy of storage estimates because of the small size of text and indices relative to the size of compressed scanned images. If some of the images will be on optical media, but all of the indices and OCR text will be on magnetic media, then the additional accuracy provided by the following section may be useful.

OCR for full text storage produces the largest indices. At 5 thousand bytes per 50 thousand byte raster scanned page image, the OCR text takes up ten percent of the storage in a system. An additional 2 to 10 thousand bytes of are required for the actual index to the OCR text, the full text index, depending on the indexing software used.

Key word and database indices rarely have more than one hundred characters per document. For one page documents, 500 bytes is one percent of the document page image size, so the database entry is less than one-fifth of one percent of the page image size, and can be ignored for most systems.

As with scanned images, the size computer generated text files can be very accurately estimated by measuring the first one percent of the documents processed when the system goes into operation.

F.2.2 Scanned Engineering Drawings / Large Format Documents

When folded, blueprints fit in a file drawer and have a thickness equal to the number of letter size documents that would cover the blueprints. For an E size drawing, this is 16 letter size documents (an E size drawing folded to fit in a letter size file folder has 16 layers of paper). Because of this relationship, the 50 KB estimate for a scanned page can be used to estimate that an E size drawing would require 800 KB ($16 \times 50,000$ bytes) of storage.

Engineering drawing sizes:

- A size (ISO/metric A4 size has approximately the same dimensions and surface area) = $8\frac{1}{2}'' \times 11''$ (or $9'' \times 12''$; equivalent to 1 letter size page)
- B size (A3) = $11'' \times 17''$ (or $12'' \times 18''$; equivalent to 2 letter size pages)
- C size (A2) = $22'' \times 17''$ (or $24'' \times 18''$; equivalent to 4 letter size pages)
- D size (A1) = $34'' \times 22''$ (or $36'' \times 24''$; equivalent to 8 letter size pages)
- E size (A0) = $44'' \times 34''$ (or $48'' \times 36''$; equivalent to 16 letter size pages)
- F size = $28'' \times 40''$
- G size = $11''$ roll \times $22\frac{1}{2}''$ to $90''$ long
- H size = $28''$ roll \times $44''$ to $143''$ long
- J size = $34''$ roll \times $56''$ to $176''$ long
- K size = $40''$ roll \times $56''$ to $143''$ long

A double truck (center fold) full broadsheet newspaper is $24\text{ in} \times 36\text{ in}$, equivalent to an old D size drawing.

1 E size drawing (1-bit black and white) = 16 letter size pages (US letter or A4) = 800 KB
1 D size drawing (1-bit black and white) = 8 letter size pages = 400 KB
1 C size drawing (1-bit black and white) = 4 letter size pages = 200 KB
1 B size drawing (1-bit black and white) = 2 letter size pages = 100 KB
1 A size drawing (1-bit black and white) = 1 letter size page = 50 KB
1 E size drawing (color) = 16 letter size pages (US letter or A4) = 1,600 KB
1 D size drawing (color) = 8 letter size pages = 800 KB
1 C size drawing (color) = 4 letter size pages = 400 KB
1 B size drawing (color) = 2 letter size pages = 200 KB
1 A size drawing (color) = 1 letter size page = 100 KB

F.2.3 Scanned Microforms

In many record series, each microfiche contain only a few images because each fiche represents a single record in the series (*e.g.* one fiche per person in a personnel record series). In this case filming breaks on records, rather than being continuous. To a lesser extent this is also true for roll film. In these cases, the amount of storage required depends on the number of images on the film, not the number of microfiche or the number of rolls of film. A full, standard 24× microfiche has 7 rows of 14 letter (or A4) size images, for a total of 98 images.

Scanned aperture card images require the same storage as the document or drawing in the aperture. This is true for any microform. The metric A6 size is equivalent to 105 mm microfiche. Aperture cards are the microform most commonly used to store blueprints. Aperture cards are punch cards that have a hole (or aperture) cut into them that holds one 35 mm slide which reproduces one blueprint sheet in most cases. An image scanned from a blueprint's image in an aperture card requires the same amount of storage needed for an image scanned from the original full size blueprint.

If you have 200 foot rolls of microfilm instead of the listed 100 foot rolls, simply multiply the number of pages by 2. A 100 ft. roll of 16 mm microfilm of 24× images actually has closer to 2,400 images rather than the listed 2,500 images. This slightly overstates the digital storage requirements, making it more conservative. This estimate must be adjusted slightly if used conversely, because it overstates the capacity of the microfilm, making slightly over-optimistic for estimating the number of rolls required for a project.

1 roll 16 mm microfilm (100 ft, ~30 m) (24× reduction) = 2,500 letter pages = 1 file drawer = 125 MB
1 roll 35 mm microfilm (100 ft) (12× reduction, open spacing, normal scan) = 1,000 letter pages = 50 MB
1 microfiche (105 mm film) (24× reduction) = 100 letter pages = 5 MB
1 roll color 16 mm microfilm (100 ft) (24× reduction) = 2,500 letter pages = 250 MB
1 roll color 35 mm microfilm (100 ft) (12× reduction, open spacing, normal scan) = 100 MB
1 color microfiche image (105 mm film) = 10 MB

F.2.4 Scanned Miscellaneous Documents

A “long” credit card receipt is functionally the same size as an 80 column punch card, which was based on the older 90 column, round hole, punched card, which in 1890 was based on the size of the old US dollar bill (before 1929). US dollar bills are now 6.14 × 2.61 inches (~156 × ~66 mm), before 1929, US dollar bills were 7.4218 × 3.125 inches (~189 × ~79 mm).

1 check (2 sided) (remittance) = 50 KB per item (less if no patterns are present)	
1 check (1 sided) (remittance) = 25 KB per item (less if no patterns are present)	
1 "long" credit card receipt (3¼" × 7⅞", 2-sided) (remittance) = 35 KB	
1 "short" credit card receipt (3¼" × 5", 2 sided) = 25 KB	
1 library book = 10 MB 50 library books = 500 MB = 1 CD 100 library books = 1 GB	1 library book (color) = 20 MB 50 library books (color) = 1 GB = 2 CDs

F.2.5 COLD / COOL

When documents are imported directly in the form they were created in, they require much less storage space. Because COLD (Computer Output to Laser Disc) pages come from mainframe computers, their formatting is very simple and thus they require even less storage space than word processor pages. COLD or COOL (Computer Output On-Line) output is compressed page and includes index. All Points Addressable (APA) pages include IBM AFP (Advanced Function Printing) and Xerox Metafont.

1 COLD/COOL letter size page = 2 KB
1 COLD/COOL 11 × 14 inch fanfolded greenbar computer sheet = 4 KB
1 COLD/COOL APA page = 10 KB

F.2.6 Digitized Multimedia Formats

Multimedia documents exist in compressed digital form, and the listing shows average sizes for these documents as well. The DVD (commonly Digital Video Disc) multimedia format standards will provide a stable foundation for working with these types of documents.

The size of compressed file depends on the resolution (dots per inch, or dpi) and the detail (information) in the photograph. The detail in a photograph is dependent on the size of the negative and the quality of the film, camera, and lens (the detail is not related to print size unless the print is smaller than the negative). The resolution of the scan should be chosen to match the detail of the photograph. For most cameras, films, and formats 35mm and smaller, the 5 MB Photo CD format (3072 × 2048 pixels) captures all the information in the image. Note that this is in dots per image rather than dots per inch.

1 hour of compressed color video = 2 GB (DVD, MPEG 2) (image quality dependent)
1 hour of audio = 10 MB (dictation, answering machine) to 500 MB (a CD holds 74 minutes of music)
1 color picture = 10 KB (thumbnail) to 5 MB (for each of 100 photos on a 500 MB photo CD)

F.2.7 Medical Records

Wavlet compression, lossless mode, has FDA 510(k) approval. Twelve bits per pixel provide 4,096 shades of gray. Images of 150 dpi and 12 bit grayscale are recommended by the American College of Radiology for primary reads. For secondary reads, wavlet compression, lossy mode, has FDA 510(k) approval. X-rays originally recorded digitally rather than on film provide a color depth of 16 bits per pixel or 65,536 shades of gray per pixel. More shades of gray allow doctors to see very fine variations in the health of tissues, increasing the early detection of disease.

1 X-ray (14 × 17 inches, 150 dpi, 12 bits, lossless wavlet compressed)= 1 MB 1 X-ray (14 × 17 inches, 150 dpi, 12 bit grayscale, lossy wavlet compressed) = 200 KB

F.2.8 Trees and Paper Requirements

As an ancillary note, each full file cabinet represents one pulp tree. Pulp trees are grown fast for paper or are trees culled from among the trees that will be allowed to grow to a larger size for lumber.

1 pulp tree (loblolly pine 8" (200 mm) diameter, 50 ft. (15 m tall, 20 years old) = .1 cord of wood (.2 cu. m of wood) = 10,000 pages = 1 file cabinet = 4 boxes/drawers = .5 GB = 1 CD-ROM
10 pulp trees (loblolly pine 8" (200 mm) diameter, 50 ft. (15 m) tall, 20 years old) = 1 lumber tree (20" (500 mm) diameter, 110 ft (35 m) tall, 50 years old) = 1 cord of wood (2 cu. m of wood) = 100,000 pages = 10 file cabinets = 40 file boxes/drawers = 5 GB = 10 CD-ROMs

Appendix G. Draft Interview Form



**US Army Corps
of Engineers®**

Automated Document Conversion Strategy (ADCS)

Site Interview Document

Army Sponsor/Point of Contact

USACE Program Manager

INTRODUCTION, PURPOSE AND SCOPE

Introduction

This worksheet relates to a study of data conversion needs for the U.S. Army Corps of Engineers, under the direction of the COR and the USACE ADCS Program Manager, to determine the Corps of Engineers' requirements for drawing conversion. Topics to be covered in the study will include at a minimum, the type of facility/structure represented by the drawings with emphasis on the probable future need for the documents; the file formats required (raster vs. vector vs. text, as well as the specifics of file formats) for the converted drawings; the details of the format and quality of the source document; the accuracy requirements of the converted documents; for conversions to CAD files, the degree of necessary adherence to CAD Standards and the specifics of those requirements; the variations in needs of different Corps Districts based on the differences in their procedures, file formats, and military customer requirements; an estimate of the quantity of each type of document to be converted to each format/accuracy category. It is anticipated that interviews will be conducted at a number of sites as a representative cross section of the Corps.

Purpose

To determine document conversion requirements.

Scope

- Details about the interview
- How long it will take
- Who needs to be interviewed
- Results end up in the Data Conversion Study and DCRDs

Site Interview Participants:

Site Location:	
Date(s) of Visit:	
Interviewer(s):	

Site POC:	
Title:	
Phone Number:	
Email Address:	

Interviewee:	
Title:	
Phone Number:	
Email Address:	

Interviewee:	
Title:	
Phone Number:	
Email Address:	

Interviewee:	
Title:	
Phone Number:	
Email Address:	

Interviewee:	
Title:	
Phone Number:	
Email Address:	

Interview Questions:

1. To what extent do engineering drawings still exist in a hard-copy format such as paper, Mylar or microfiche? (total number)

Format	0	5K ^{+/-}	10K ^{+/-}	25K ^{+/-}	50K ^{+/-}	75K ^{+/-}	100K ^{+/-}
Paper							
Mylar							
Aperture Cards							
Microfiche							
TIFF							
CALS Type 1							
Raster							
Outdated CAD							
Outdated File Format							

CAD format(s)

File Format(s)

2. Describe the type of facilities/structures represented by the drawings with emphasis on the probable future need for the documents.

3. Describe how the information stored in the engineering drawings might be used today (if it were accessible) or in the future?

4. Has someone appraised the drawings to estimate what percentage might need to be converted in order to use them in existing or future business processes?

☐ Yes ☐ No

Format	0	5% ^{+/-}	10% ^{+/-}	15% ^{+/-}	20% ^{+/-}	25% ^{+/-}	More
Paper							
Mylar							
Aperture Cards							
Microfiche							
TIFF							
CALS Type 1							
Raster							
Outdated CAD							
Outdated File Format							

5. Are there initiatives already underway to convert the drawings?

☐ Yes ☐ No

6. Are there conversion initiatives planned for the coming 2 months?

☐ Yes ☐ No

7. What disciplines do the drawings represent?

Discipline	Paper	Mylar	Aperture	Microfiche	TIFF	CALS
Architectural						
Telecommunications						
Civil/Site						
Civil/Works						
Electrical						
Fire Protection & Suppression						
General						
Geotechnical						
Interior Design						
Landscaping Architecture						
Mechanical						
Plumbing						
Equipment (Security)						
Structural						
Survey & Mapping						
Utilities						

Discipline	Raster	Outdated CAD	Outdated File Format
Architectural			
Telecommunications			
Civil/Site			
Civil/Works			
Electrical			
Fire Protection & Suppression			
General			
Geotechnical			
Interior Design			
Landscaping Architecture			
Mechanical			
Plumbing			
Equipment (Security)			
Structural			
Survey & Mapping			
Utilities			

8. Describe the environment the drawings are stored in.

9. Describe the condition of the drawings.

10. Is there a paper record, computerized index or drawing management system containing the pertinent title block information which describes the contents of the drawings?

11. Is there an effort under way now to record the title block information?

12. Is there a need, beyond transferring some of the drawings in to a raster format? Say an intelligent CAD format for example. If so, which CAD format would be most useful?

13. What number of the drawings do you think would require being dimensionally accurate?

14. Does the information contained in the drawings need to conform to some CAD Standard?

15. Is the CAD Standard consistent with the A/E/C CAD Standards? If no, is the CAD standard used documented?

16. Is that CAD Standard documented, in place and accepted in today's workflow?

17. Is the format of the sources documents recorded somewhere?

18. What output formats would work best for your workflows and satisfy the requirements of your customers/contractors?

Input Types	Output Types		Adherence to Standards	Accuracy Required
Paper	<input type="checkbox"/> Raster	<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	
Mylar	<input type="checkbox"/> Raster	<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	
Raster		<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	
Aperture Cards	<input type="checkbox"/> Raster	<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	
Microfiche	<input type="checkbox"/> Raster	<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	
TIFF		<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	
CALS Type 1		<input type="checkbox"/> CAD	<input type="checkbox"/> YES <input type="checkbox"/> NO	

(We need to understand the variations in needs between the different Corps Districts based on their procedures, file formats, and military customer requirements.)

19. Are there legacy drawings stored on magnetic media?

If so, what are their file formats?

Do different groups have different file formats? ☐ Yes ☐ No

What are they?

How are the files stored (in an EDMS, on CDs, magnetic tape, etc.)?

20. What types of magnetic media are involved?

Metadata Related Questions:

21. What metadata has been collected for documents that have already been converted?

22. What metadata is anticipated to be needed for inclusion in an EDMS?

23. Is an EDMS currently installed or in the process of being installed?

24. If yes, what is the system?

25. What is the purpose of the EDMS and who are its anticipated users?

26. What metadata is being collected and indexed?

27. What metadata is anticipated to be needed for Records Management purposes?

28. What information is someone who is searching for a document likely to have available to aid in locating the document?

In a year?

In 5 years?

In 10 years?

29. Is your planned strategy to have multiple repositories for converted documents, or a single one? Why?

30. What is driving your need to convert documents to electronic formats?

31. What is your time requirement for conversions and is the conversion funded?

32. Is your intent to have a single flat file format for all metadata regardless of (business) document type, or different schemas for different types?

POINTS OF CONTACT

Appendix H. Sample DCRD